# Phenotypic Innovation, Robustness and Recombination in Complex Metabolic Systems

**Dissertation**

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
**Sayed Rzgar Hosseini**
dem
Iran

**Promotionskommission:**
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Olivier Martin
Prof. Dr. Frédéric Guillaume

Zürich, 2018

**I dedicate all my achievements**

*To my late beloved mother for everything, Words are not sufficient to express my gratitude to you…*

*To my wonderful brothers, Vafa, Hiva and Kamran who are always the greatest sources of inspiration and support in my life, and*

*To my great father who has taught me in practice how to remain persistent and not to give up on my dreams.*

# Abstract

How novel phenotypes emerge in biological systems is a fundamental question of evolutionary biology that has been left poorly answered due primarily to the inherent complexity of establishing genotype-phenotype maps in complex biological systems. Thanks to the availability of comprehensive genomic and biochemical information about metabolic systems and efficient computational methods, mainly the experimentally validated method of flux balance analysis, I could systematically establish a quantitative framework to gain unprecedented insights into the causes and origins of phenotypic innovation in metabolic systems. My analyses revealed the power of recombination in bringing forth novel phenotypes in genome scale metabolisms. Because innovation and robustness are highly interrelated concepts, I also systematically analyzed the robustness of bacterial genomes to pervasive large-scale gene deletions. My analyses revealed an organization of bacterial genomes that ensures substantially higher robustness to such destructive events than expected by chance. I followed these observations by rigorous analyses to identify the evolutionary mechanisms that can create such an organization. In this quantitative framework, I could also analyze constraints and contingencies in complex metabolic systems to show how predictable phenotypic innovation is. Furthermore, by establishing an exhaustive genotype-phenotype map in central carbon metabolism, within the framework of genotype networks, I have shown how phenotypic innovation is facilitated during evolution, and to what extent non-adaptive mechanisms such as exaptation can explain the origin of novel phenotypes in metabolic systems.

# Zusammenfassung

Ein grosses Teilgebiet der Evolutionsbiologie befasst sich mit der fundamentalen Frage der Entstehung neuartiger Phänotypen in biologischen Systemen. Das Erstellen von Genotyp-Phänotyp-Karten in biologischen Systemen ist sehr komplex und daher wurde die obige Frage bis jetzt nicht zufriedenstellend beantwortet. Dank der Verfügbarkeit umfassender genomischer und biochemischer Informationen über Stoffwechselsysteme und aufgrund effizienter Berechnungsmethoden wie der experimentell validierten „flux balance analysis“ (FBA), konnte ich dieser Frage quantitativ und systematisch nachgehen. Meine Analysen erlauben beispiellose Einblicke in die Ursachen und Ursprünge phänotypischer Innovationen in Stoffwechselsystemen. Sie demonstrieren zum Beispiel die Fähigkeit der Rekombination, neue metabolische Phänotypen hervorzubringen. Da Innovation und Robustheit in hohem Maße miteinander zusammenhängen, analysierte ich auch systematisch die Robustheit von bakteriellen Genomen gegenüber Deletionen eines oder mehrerer Gene, welche in der Genomevolution häufig vorkommen. Weiters konnte ich durch meine Analysen eine Organisation bakterieller Genome feststellen, welche eine wesentlich höhere Robustheit gegenüber solchen destruktiven Ereignissen aufweisen als der Zufall erwarten lässt. Mittels gründlicher Analysen identifizierte ich die evolutionären Mechanismen, welche eine solche Organisation zustande bringen. In diesem quantitativen Rahmen war es mir auch möglich, evolutionäre Einschränkungen („constraints“) und Kontingenzen in komplexen metabolischen Systemen zu analysieren und dadurch die Vorhersehbarkeit phänotypischer Innovation aufzuzeigen. Des weiterem habe ich im Rahmen von Genotyp-Netzwerken eine umfassende Genotyp-Phänotyp-Karte des zentralen Kohlenstoffmetabolismus erstellt. Dadurch konnte ich zeigen, wie phänotypische Innovation während der Evolution erleichtert wird und inwiefern nichtadaptive Mechanismen wie die Exaptation den Ursprung von neuen Phänotypen in metabolischen Systemen erklären können.

# Contents:

# Chapter 1: Introduction

## 1.1. Phenotypic innovation

Theodosius Dobzhansky indeed did not exaggerate when he stated, "Nothing in biology makes sense except in the light of evolution", because the theory of evolution by natural selection proposed by Charles Darwin powerfully unifies all of life's marvelous diversity. Nevertheless, there is an important aspect of life that Darwin's theory is unable to explain. Although his theory of evolution perfectly explains how already-existing variation spread by natural selection, it is completely silent on how new variants come into existence in the first place. The problem is that natural selection per se cannot innovate, because it is not a creative force; it merely selects what is already there [1]. Hugo de Vries, the Dutch geneticist who rediscovered Mendelian laws of heredity, beautifully expressed this problem when he said that "Natural selection may explain the *survival* of the fittest, but it cannot explain the *arrival* of the fittest".

How do the new variants that natural selection needs arise? In other words, how does nature innovate? A common sense answer would be that new variants arise randomly, by chance. However, Darwin was himself aware of the fact that chance alone can explain nothing, when at the beginning of the chapter on laws of variation in the *Origin* [2] he wrote that:

" I have hitherto sometimes spoken as if the variations … had been due to chance. This, of course, is a wholly incorrect expression, but it serves to acknowledge plainly our ignorance of the cause of each particular variation. "

More than 150 years after Darwin's theory, we still do not have a systematic understanding of the origin of new variants, mostly because of the following two reasons:

First, we have so far focused mostly on *genotypes*; that is we have focused our attention mostly on already-existing variation at the genotypic level and how it spreads. After rediscovery of the Mendelian laws of heredity, and after the introduction of the concept of the gene, the theory of evolution was extensively used in quantitative and mathematical studies to model the dynamics of genetic variation. This resulted in the emergence of a discipline called population genetics, which views

a population as a collective pool of genes instead of organisms. Nourished by the mathematical and statistical insights of great intellectuals like Sewall Wright, J. B. S. Haldane, and Ronald Fisher, population genetics became a mature field able to quantify how genetic variants spread. Advances in molecular biology, genome science and high throughput sequencing technologies permitted the application of evolutionary theory to genome evolution and paved the way for the birth of population genomics. However, even in population genomics the central theme remains intact; we can quantify at an unprecedented level how genetic variations spread, but despite all these technological advances, we still do not have a systematic understanding on how new variation emerges in the first place.

Second, organismal *phenotypes* are highly complex, which renders the prediction of phenotypes from genotypes a very challenging task. Despite the efforts of thousands of biologists over many decades we still struggle to fully understand the phenotype of even the simplest organisms, because it is not easy to fully understand how genes help shape this phenotype. Population and quantitative genetics gradually included more complex phenotypes by allowing genes contribute in non-linear ways to a phenotype. They also study multivariate phenotypes that are represented as vectors not as a single scalar. Nevertheless, these representations cannot capture the entire complexity of phenotypes such as the fold of a protein. This phenotype, which includes the molecular motions of a protein's amino acids, is so complex that we cannot compute it from information in the genotype. Moreover, most of the complex tasks inside a cell are performed by multiple proteins, which interact with each other in a coordinated way. In other words, underlying complex organismal phenotypes are networks of interacting molecules that form signaling, regulatory and metabolic networks. Hence, to predict a given phenotype we need to identify the underlying genotypic players and to understand how exactly they interact with each other. Systems biology has made significant progress in identifying these biological networks. For example, it has established interaction maps, which show who interacts with whom. However, merely listing the molecules and their interaction partners does not help us to understand complex phenotypes. We need to understand exactly how these molecular players interact with each other to form a whole phenotype. This has become possible to some extent through mathematical models that aim to describe how the concentrations and activities of molecules change over time. These models

need parameters, but these parameters are not fully known for all the processes, pathways and networks existing in a given organism. Moreover, the resulting mathematical equations are not simple to solve analytically.

Therefore, in order to predict complex phenotypes from genotypes, we must be able to address these problems by choosing right modeling approaches. The models should be as simple as possible, and at the same time they must be able to capture the essential characteristics of the system, as Albert Einstein said "everything should be as simple as possible, but not simpler". Population geneticists used simplified models to understand the evolution of genes and genotypes, and they mostly ignored the complexity of organismal phenotypes. These simplifications were crucial for population geneticists to understand natural selection in action. However, to understand the origin of novel phenotypes and innovability, we cannot ignore the complexity of organismal phenotypes, but we must embrace it. In other words, we must gain sufficient knowledge on how genotypes are mapped to their corresponding phenotypes. Thus, to approach the problem of innovation, we need to focus on biological systems for which the relationship between genotype and phenotype is well understood, and we must choose the right computational approaches to tackle their inherent complexity.

Although it is inevitably difficult, let me define phenotypic innovation more precisely with the help of some examples. Consider the phenotypic differences between a complex organism and that of the simplest life forms. Each of these differences can be considered as an adaptive solution that emerged as a response to a particular challenge faced by an organism during its evolutionary history. For example, photosynthesis can be seen as an innovative response to convert light energy to chemical energy in order to fuel various essential cellular activities. Likewise, the requirement for moving from one place to another has led to the emergence of innovative responses as diverse as whale's tail fluke and the bacterial flagellum. In general, we can define phenotypic innovation as "*a new feature that endows its bearer with qualitatively new, often game-changing abilities* [3]". The phenotypic changes during major transformations in the history of life are magnificent examples of innovation. They include the emergence of plants with flowers, animals with hard skeleton, birds and insects with wings, multicellularity, organisms living in groups, teeth to digest hard foodstuffs, vascular systems of plants and nervous systems in

animals and etc. [3]. Another broad class of examples, which will be the focus of my thesis, are the metabolic innovations that emerge when an organism acquires the ability to use alternative sources of energy or produce novel chemical substances.

## 1.2. Metabolic systems and innovation

To systematically study phenotypic innovation in my Ph.D. research, I have focused exclusively on metabolic systems. I did so for three major reasons: First, metabolism is one of the most fundamental biological systems required for sustaining life. Second, metabolism is the source of many well-documented innovations, and especially in prokaryotic organisms. Third, we have efficient and experimentally validated computational methods that can systematically establish genotype-phenotype maps in metabolic systems.

Metabolism and replication are the two fundamental requirements for sustaining life. A minimal life form requires a metabolism, a network of chemical reactions that extracts energy and creates molecular building blocks of life. The reactions in this network are accelerated (i.e. catalyzed) by enzymes, which are encoded by metabolic genes. Broadly speaking, a metabolism encompasses two complementary kinds of chemical transformations. The first extracts energy from energy-rich molecules such as glucose, and the second uses the extracted energy to transform nutrient molecules to molecular building blocks such as amino acids in proteins, DNA nucleotides, RNA nucleotides and lipids. In addition, metabolism manages body's waste, for example by converting toxic molecules to harmless ones.

Besides its fundamental roles in life, metabolism is the source of countless well-documented innovations. Especially, bacterial species are astonishing metabolic innovators. They can metabolize a bewildering variety of natural, synthetic, and even toxic molecules. These novel abilities help bacterial species to survive in unusual environments and let them explore or invade new habitats. For example, microbes have been reported to thrive on toxic industrial substances such as polychlorinated biphenyls [4], chlorobenzenes [5], organic solvents, and synthetic pesticides like pentachlorophenol [6,7]. One study on microbial isolates from pristine soils showed that some microbes not only are not killed by several antibiotics including ciprofloxacin, a fully synthetic compound, but also are able to use those antibiotics as food [8].

Another way by which microbes innovate is by synthesizing novel compounds. For example, halophilic bacteria can survive on saturating salt concentrations of 30% by synthesizing compatible solutes such as ectoine or glycine betaine, which are able to stabilize proteins and neutralize the high osmotic pressure caused by high salt concentrations [9–11]. Another example is the synthesis of light-harvesting chlorophyll pigments, which was a key innovation step in the evolution of photosynthesis [12,13].

Although metabolic innovation is mostly documented in the prokaryotic world, we can also find metabolic innovations in higher animals. A typical one is the urea cycle in land living organisms. Ammonia, which is a toxic product of animal metabolism is converted to less toxic compounds in the urea cycle. Importantly, the urea cycle exemplifies the combinatorial nature of innovation: The individual reactions are not necessarily new, but it is their combination that is new. The urea cycle is formed by combining a set of four reactions involved in arginine biosynthesis with arginase, a reaction involved in arginase degradation [14]. These reactions are not new, as they all exist in both prokaryotes and eukaryotes, but their novel combination allows the conversion of ammonia to urea.

Plants are also particularly ”talented“ metabolic innovators. They can produce an amazing diversity of chemical compounds known as secondary metabolites [15]. These compounds confer a wide variety of selective advantages. For example, floral scent pigments increase fertilization rates by attracting insect pollinators [16,17]. Toxic secondary metabolites can play defensive roles by repelling pathogens and herbivores [18,19]. Specific chemical compounds found in fruits can prevent spoilage, and finally, specific metabolites can help plants grow in harsh conditions such as high salt concentrations or arid environments [20,21].

Fortunately, thanks to the efforts of many experimental labs over decades, we have accumulated systematic knowledge about the biochemical reactions of metabolism in many different species from bacteria to human. Moreover, technological advances permitting genome sequencing of many different species have provided us with an unprecedented opportunity to reconstruct genome-scale metabolic networks from genomic data [22,23]. Comprehensive information about thousands of reactions with their corresponding enzymes, enzyme-coding genes, and gene-reaction association rules are stored in biological databases such as KEGG (Kyoto Encyclopedia of Genes

and Genomes) [24–26], BIGG (Biochemical and Genetic and Genomic database) [27], and SEED (a database and infrastructure for comparative genomics) [28].

The final reason that makes metabolism an ideal system for studying innovation is the availability of efficient, well established and experimentally validated computational methods for predicting metabolic phenotype from metabolic genotype. I discuss these computational approaches in more detail in the following section.

## 1.3. Computational modeling of metabolic systems

### *a) Kinetic modeling*

Classical biochemical studies were based on small-scale analyses of metabolic systems including a handful of reactions or a linear sequence of reactions [29], and modeling the behavior of such small systems was based on kinetic models using ordinary differential equations [30]. In order for these equations to accurately predict the concentrations of metabolites over time, they need precise experimentally determined parameters. For example, in the following Michaelis-Menten equation of enzyme kinetics [31]

$$\frac{d[\mathrm{P}]}{dt} = \frac{V_{max}\ \ [\mathrm{S}]}{K_m + [\mathrm{S}]} \quad (1)$$

The concentration of product ([P]) over time is described by the concentration of a substrate ([S]) and two parameters that are required to be determined experimentally as precisely as possible. These are $V_{max}$, the maximum catalytic rate achievable by the enzyme at saturating substrate concentration, and $K_m$ (the Michaelis constant), which is the substrate concentration at which the reaction rate is half of $V_{max}$. Moreover, in such kinetic models, all regulatory information, including allosteric interactions, which are defined as regulation of enzyme activity by effector molecules outside of the enzyme's active site, should be considered. These kind of kinetic models are not easily scalable for analyzing genome-scale metabolic networks including thousands of reactions and enzymes, for which most parameters and regulatory interactions are unknown. Therefore, alternative modeling approaches, which sacrifice precision in order to scale up to large-scale systems, are required.

***b) Constraint-based modeling***

To circumvent the need for experimentally measured kinetic parameters and the lack of sufficient knowledge about the underlying regulatory interactions, a coarse grained approach based on constraint-based modeling called flux balance analysis (FBA) is often used to model genome-scale metabolic systems. FBA is based on a steady state assumption, that is, the concentrations of metabolites in the cell are stationary and do not change over time. Moreover, FBA only requires stoichiometric information about reactions. Thus, it is completely independent of kinetic parameters. Finally, in this method, regulatory information about enzymatic reactions is completely neglected. Nevertheless, FBA is able to predict system level qualitative phenotypes accurately, because its predictions are consistent with experimental studies [32]. For example, FBA is able to accurately predict whether a given metabolism is able to produce all essential biomass precursors from a given nutrient or not, even though it may not provide accurate predictions for the flux of all individual reactions.

FBA predicts the metabolic flux of each reaction based on the stoichiometric coefficients of the metabolites participating in the reactions of a given metabolic network. Stoichiometric coefficients are stored in a stoichiometric matrix *S*, which is of dimension $m \times n$, where $m$ and $n$, respectively, denote the number of metabolites and the number of reactions in a metabolic network. The flux through each reaction is constrained based on the assumption that metabolite concentrations do not change, i.e., $Sv = 0$, where $v$ is the vector of metabolic fluxes $v_i$ through reaction $i$. The solutions of the equation $Sv = 0$, that is, the nullspace of matrix *S*, comprises all flux vectors that are allowable in steady state.

Thus, there may be infinitely many solutions within the null space that can satisfy the steady state assumption. Therefore, the null space has to be constrained by additional physicochemical information regarding the maximum and minimum possible flux through each reaction. FBA relies on an optimization procedure called linear programming to identify those flux vector(s) among the allowable ones that maximize an objective function $Z$ such as ATP production rate or biomass production rate that is, the rate at which metabolic compounds are converted into biomass constituents. This task can be formulated as finding a flux vector $v^*$ with the property

$v^* = max_v\, Z(v) = max_v \{ c^T v \mid Sv = 0,\ a \leq v \leq b \}$, (2)

where the vector $c$ contains a set of scalar coefficients representing the maximization criterion, and each entry $a_i$ and $b_i$ of vectors $a$ and $b$, respectively, indicate the minimally and maximally possible flux through reaction $i$. It is important to note that in our applications, the vector $c$ is a binary vector, whose elements are all zero except the element corresponding to the biomass reaction. Therefore, $v^*$ maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass [33].

An alternative constraint-based approach with widespread applications in metabolic engineering is Elementary Flux Mode Analysis (EFMA) [34]. Similar to FBA, EFMA also solves $Sv = 0$ subjected to the imposed constraints. However, in EFMA the aim is not to calculate a specific flux vector to optimize an objective function. Instead, the goal is to identify all possible flux vectors (or flux modes) belonging to the nullspace, which satisfy the following "elementary flux mode" property. The set of non-zero indices in a flux vector (i.e. a mode) are called the support of the flux mode. A flux mode (e) is called elementary (EFM), if its support is not a superset of the support of any other feasible flux mode (v) ($\mathrm{supp(e)} \not\supset \mathrm{supp(v)}$). An EFM is thus a minimal and unique set of reactions with non-zero flux in steady state [35]. If any of the reactions with non-zero flux is deleted, the EFM is not able to remain functional in steady state anymore [36]. In other words, EFMs are non-decomposable steady-state pathways through a network. The most important advantage of EFMs is that any steady state flux vector can be represented as a non-negative weighted sum of EFMs.

The full metabolic capabilities of a metabolic network can be expressed by its set of elementary flux modes [37]. EFMA is employed in the first steps of metabolic engineering to check the feasibility of production of a given metabolite from a given substrate [35]. Production is possible, if there is at least one EFM that connects the substrate and the product of interest. Moreover, EFMA is further employed in the optimization step of metabolic engineering projects by identifying the target reactions that do not participate in the desired EFMs and whose elimination can improve the production yield of desired metabolites [38,39]. Experimental studies have confirmed the power of EFMA in the construction of minimal metabolic networks with optimal production efficiency [40–44].

Exhaustive enumeration of all possible EFMs is possible only for small metabolic systems, but for larger systems, the calculation of EFMs represents a major algorithmic challenge, which limits EFMA's applicability [35]. Several methods have been introduced to tackle the computational complexity of EFMA [45–50]. The common core of all these algorithms is the same. They generate EFMs by pairwise combination of existing EFMs, followed by verification of the generated EFMs to ensure that they have not been identified before. The verification step is the major bottleneck of these algorithms. To make this step computationally more efficient, several algorithmic improvements have been introduced, including the application of binary representation [51], matrix rank [52], bit pattern trees [53], distributed memory parallelization [54], and efficient sampling-based approaches [55,56].

Because in my studies I analyze millions of large-scale metabolic networks, the computational efficiency is very important to make my studies feasible. Therefore, FBA, which is parameter-free and efficient, is the method of my choice in this dissertation. It allows me to systematically analyze many different metabolic systems required for studying metabolic innovation. I use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and I consider a metabolic network viable if it can produce all essential biomass precursors.

### *c) $C^{13}$ based metabolic flux analysis*

FBA determines the reaction fluxes based on the assumption that cellular metabolism maximizes a specific objective function. However, this optimization principle may not reflect biological reality, because cellular metabolism seems to display suboptimal performance in some biological systems [57,58]. Thus, to obtain more realistic estimates of metabolic fluxes, it can be desirable to quantify intracellular steady state fluxes within a predefined metabolic network under in vivo conditions.

This approach uses experimental information from isotope tracer experiments to infer intracellular flux distributions. Specifically, $C^{13}$-labeled substrates are fed to a growing cell population until the isotope label is distributed through the metabolic network [59]. As a function of the particular distribution of metabolic fluxes in an organism, specific labeling patterns occur in metabolic intermediates [60,61] . Data based on $C^{13}$ labeling are obtained by analytical methods such as NMR [62–64] or

mass spectrometry [65–67]. To infer intracellular fluxes from $C^{13}$ labeling patterns, one of the following two complementary procedures is used.

In the first procedure, $C^{13}$-labeled based data and extracellular fluxes are integrated into a computational model. The goal is to fit flux of metabolic reactions to experimental data. The fluxes are fitted iteratively to measured data until the difference between observed and simulated isotope spectra is minimized [61]. The second procedure relies on direct interpretation of selected labeling patterns. In this approach, a flux ratio is derived that quantifies the relative contribution of converging pathways to the formation of particular metabolites from a given NMR or mass spectrometry pattern [66,68].

Due to the complexity of flux inference procedures and the experimental difficulty associated with carbon labeling experiments, this approach cannot be applied to studying genome-scale metabolic systems. Instead, it is only applicable for small metabolic systems with 50 to 100 reactions, such as central carbon metabolism [66]. Nevertheless, $C^{13}$ metabolic flux analysis has played a major role in biotechnology and metabolic engineering to produce molecules with industrially important capabilities, such as the amino acids lysine and glutamate [69,70]. Moreover, it has been used to study the metabolism of microorganisms engineered for secondary metabolite production, such as penicillin production in *Penicillium chrysogenum* [71].

***d) Metabolic control analysis***

One aspect of metabolic systems that is not quantifiable by the approaches I described thus far is how cells regulate and control their fluxes. Our knowledge of metabolic systems remains incomplete until we completely uncover all the details of metabolism's regulation. Identifying the mechanisms controlling flux through a particular pathway and quantifying the extent of the pathway's control require alternative computational approaches. Gaining insight into these control mechanisms is particularly important for metabolic engineering purposes; for example, how to optimally manipulate fluxes to maximize the production of a given metabolite. These quantifications can be achieved to some extent through a computational method called Metabolic Control Analysis (MCA), which defines a quantitative link between pathway fluxes and the activity of a pathway's constituent enzymes [72].

MCA quantifies how variables such as reaction fluxes or metabolite concentrations depend on network parameters. In this mathematical framework, network dependent properties are encapsulated in control coefficients and MCA describes how these coefficients depend on local properties called elasticities. MCA is fundamentally a first order sensitivity analysis in the vicinity of a fixed point that is the steady state of a metabolic system. The relative steady state change in a system variable such as pathway flux (*J*) or metabolite concentration (*Sc*) in response to a relative change in a parameter (*P*) such as enzyme activity or the steady state flux of the $i^{th}$ reaction ($v_i$) is quantified by control coefficients. The two main control coefficients are the flux and concentration coefficients. The flux control coefficients $C_{v_i}^{J}$ are defined as:

$$C_{v_i}^{J} = \left(\frac{\mathrm{dJ}}{\mathrm{dP}} \quad \frac{\mathrm{P}}{\mathrm{J}}\right) / \left(\frac{\partial \mathrm{v_i}}{\partial \mathrm{P}} \quad \frac{\mathrm{P}}{\mathrm{v_i}}\right) = \frac{\mathrm{dlog\,J}}{\mathrm{dlog\,v_i}} \qquad (3)$$

And the concentration coefficients $C_{v_i}^{Sc}$ are defined as follows:

$$C_{v_i}^{Sc} = \left(\frac{\mathrm{dSc}}{\mathrm{dP}} \quad \frac{\mathrm{P}}{\mathrm{Sc}}\right) / \left(\frac{\partial \mathrm{v_i}}{\partial \mathrm{P}} \quad \frac{\mathrm{P}}{\mathrm{v_i}}\right) = \frac{\mathrm{dlog\,Sc}}{\mathrm{dlog\,v_i}} \qquad (4)$$

The control coefficients of different reactions are not independent of each other, because metabolic fluxes are system properties. Thus, their control is shared by all reactions in the system. For a metabolic pathway this is formulated as the summation theorem [73,74]:

$$\sum_i C_{v_i}^{J} = 1 \qquad (5)$$

$$\sum_i C_{v_i}^{Sc} = 0 \qquad (6)$$

It implies that when a given reaction changes its control of flux, the change is compensated by changes in the control of the flux by all other reactions.

The local response of a chemical reaction to changes in its environment, such as changes in substrate or product concentrations, is measured by elasticity coefficients ($\varepsilon_{Sc}^{i}$). The relationship between elasticity and control coefficients in a metabolic pathway is expressed by the connectivity theorem, which highlights the close relationship between the kinetic properties of individual reactions and the systems properties of a pathway [73,75]:

$$\sum_i C_i^J \varepsilon_{Sc}^i = 0 \qquad (7)$$

$$\sum_i C_i^{Sc_n} \varepsilon_{S_m}^i = 0, \;\; n \neq m \qquad (8)$$

$$\sum_i C_i^{Sc_n} \varepsilon_{Sc_m}^i = -1, \;\; n = m \qquad (9)$$

Equation 7 (i.e., the connectivity theorem for flux-control coefficients) implies that for a common metabolite *Sc*, the sum of the products of the flux-control coefficient of all steps affected by *Sc* and its elasticity coefficients towards *Sc*, is zero. For the concentration-control coefficients, the analogous equations 8 and 9 apply. Equation 8 applies to the case in which the reference metabolite ($Sc_n$) is different from the perturbed metabolite ($Sc_m$). Equation 9 applies to the case in which the reference metabolite is the same as the perturbed metabolite. The connectivity theorems describe how perturbations on metabolites of a pathway propagate through the chain of enzymes. The *local* (kinetic) properties of each enzyme effectively propagate the perturbations to and from its immediate neighbors.

Finally, by combining the summation theorem with the connectivity theorem, we can obtain closed expressions that relate control coefficients to elasticity coefficients. Based on explicit expressions for control coefficients we can quantify the control mechanisms of a metabolic system. For example, we can determine whether the flux through a given metabolic pathway is controlled predominantly by one single reaction (traditionally referred to as a rate limiting step) or whether the control is distributed among all the reactions in the pathway. This invaluable information can be applied to identify the regulatory mechanisms in metabolic pathways and can help to optimally engineer metabolic pathways for the production of metabolites of industrial or medical importance [76–78].

In my analyses, my aim is not to discover the regulatory mechanisms in metabolic systems, and I am not concerned about the properties of individual reactions in metabolism. Therefore, I do not utilize MCA in my studies. I use FBA, which neglects the regulatory mechanisms in metabolic systems. However, for the qualitative phenotypes that I am concerned with, namely the viability or inviability on a given set of carbon sources, this simplification is not too unrealistic, especially if we consider the fact that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments [79–81].

***e) Markov-Chain Monte Carlo Sampling***

Systematic understanding of phenotypic innovation is only achievable through systematic genotype-phenotype maps, where thousands and millions of genotypes with their corresponding phenotypes are analyzed. Therefore, for most of the research described here, my analyses requires going beyond any known biological system and analyzing thousands and millions of potential metabolisms with new combination of reactions. To this end, I utilize the concepts of a reaction universe and a metabolic genotype space in my analyses. The reaction universe comprises all N reactions known to exist in some prokaryotic species. If we represent a given potential metabolism (i.e. a metabolic genotype) as a binary vector whose length is identical to the number of reactions in the reaction universe (N), the total number of possible genotypes is $2^N$. These comprise the metabolic genotype space.

In part of my studies, where I focus on central carbon metabolism, the reaction universe only has 51 internal reactions [23,82], which allows establishing an exhaustive genotype-phenotype map. Having access to such a map, I was able to study innovation in the framework of genotype networks [83,84]. All genotypes with the same phenotype are arranged in a genotype network, where two genotypes are connected by an edge if they differ from each other by the smallest genotypic difference, which corresponds to a single reaction in metabolic systems. This representation allows us to utilize the concepts and methods developed in graph and network theory [85] to facilitate analysis of the evolution of metabolic systems.

Unfortunately, for genome-scale metabolic systems such an exhaustive approach is not possible because the reaction universe contains more than 5000 reactions. Thus, the genotype space is astronomically large ($2^{5000}$ genotypes). Therefore, analyzing genotype-phenotype relationship in genome-scale metabolic systems is possible only through sampling approaches. In other words, I needed to sample thousands of genotypes with a given phenotype from this vast genotype space. To this end, I used the Markov Chain Monte Carlo (MCMC) method [86], a widely used approach for sampling from large and high dimensional spaces, that have also been applied to study metabolic systems [87,88]. In each step of the MCMC algorithm, a genotype is generated from a previous genotype using a probabilistic transition rule. At each transition, a small modification to the current genotype is proposed, and the modified genotype is accepted as the next genotype if the phenotype does not change,

otherwise the modification proposal is rejected (Figure 1). Therefore, this algorithm generates a sequence of genotypes with the same phenotype. To sample metabolic genotypes, the modification that I introduce at each transition step is a reaction swap that adds a randomly chosen reaction from the reaction universe to the current genotype, and deletes a randomly chosen reaction from the current genotype. The MCMC algorithm is thus a random walk in the subspace of the genotype space containing all genotypes with the same phenotype. It requires an initial genotype and in most of my analysis I used the genotype of a well-studied organism like *E. coli*. Previous studies have shown that in order to ensure that the sampled genotypes are not biased by the initial genotype, at least 3000 MCMC steps are needed before saving the first chosen genotype [87,88]. Moreover, due to strong autocorrelations between successive genotypes, it is advisable to only save every $1000^{th}$ genotype in the sample. Thus, to sample 1000 genotypes, running an MCMC algorithm with $10^6$ steps (successful reaction swaps) will be needed.

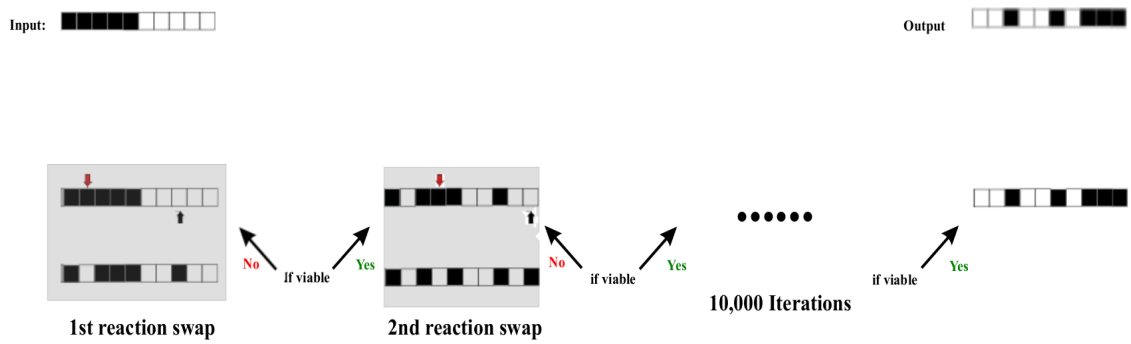



**Figure 1: MCMC Sampling of random viable genotypes.** Genotypes are represented as arrays of black and white boxes, each of which correspond to a given reaction in the reaction universe. Black boxes indicate that the corresponding reaction is present in a genotype, and white boxes show the absence of the corresponding reaction. The input of the MCMC algorithm is an initial genotype (a known genotype like that of *E. coli*). The initial genotype is subjected to 10,000 steps of phenotype-preserving reaction swaps (gray rectangles), in which a randomly chosen present reaction is removed (indicated by a red arrow) and a randomly chosen absent reaction is added (indicated by a black arrow). Then the phenotype of the new genotype is computed by FBA to ascertain viability. If the new genotype is viable, the reaction swap is accepted and the new genotype becomes the input of the next reaction swap, otherwise the reaction swap is not accepted. The viable output of the 10,000-th reaction swap, which is a viable random genotype, is the output of the MCMC algorithm.

## 1.4. Prokaryotic recombination

As I briefly mentioned in section 1.2, innovation in metabolic systems has a combinatorial nature, that is, new combinations of already existing reactions can lead to the emergence of novel phenotypes. As an example, the ability to digest the synthetic pesticide pentachlorophenol by *Sphingomonas chlorophenolica* is caused by the formation of a new metabolic pathway comprising four reactions catalyzed by enzymes that process natural chlorinated chemicals and by an enzyme involved in tyrosine metabolism [7]. Similarly, *Arthrobacter aurescens* can thrive on the man-made pesticide atrazine as a sole carbon and nitrogen source by assembling six reactions that exist in other organisms into a new pathway [89].

Recombination as a pervasive genetic force, which helps to form new combinations of already existing components might therefore be a strong genetic force behind phenotypic innovations. There is compelling evidence attesting to the power of recombination in the origin of novel phenotypes based on DNA shuffling, a widely used experimental technique to engineer novel biological systems [90]. This technique generates recombinant DNA molecules consisting of multiple recombinant fragments of several parental DNA molecules [91]. For example, in an experiment to create antibiotic resistant cephalosporinases, DNA shuffling yielded a 270 fold increase in resistance to moxalactam, as compared to an only 8 fold increase in resistance by mutagenesis [92]. Moreover, for improving production of the antibiotic tylosin in the bacterium *Streptomyces fradiae*, genome-scale recombination through DNA shuffling is 20 times more effective than random mutagenesis [93].

Although prokaryotes are haploid and lack the meiotic recombination of eukaryotes, they can still recombine genetic material. Prokaryotic recombination is not rare and can occur much more frequently than point mutations [94–99]. Deep sequencing of bacterial populations and metagenomic sequencing have shown that homologous recombination in bacteria leads to pervasive genetic mosaicism in different bacterial and archaeal species of diverse extremophiles, mesophiles, terrestrial and aquatic bacteria, and pathogenic and non-pathogenic microbes [94–98,100–107]. Although the rate of homologous recombination decreases exponentially with genomic sequence divergence, it can still be substantial even for a nucleotide divergence of 10% or greater [100,103].

The most prominent mechanism for adding genes to prokaryotic genomes is horizontal gene transfer. It leads to addition of new reactions to metabolic networks when enzyme-coding genes are transferred. The mechanisms of horizontal gene transfer include bacterial conjugation via transferrable plasmids, transduction of DNA mediated by viruses, and transformation (i.e., the uptake of naked DNA from the environment) [108]. Horizontal gene transfer can change the organization of genomes on relatively short evolutionary time scales and its rate strongly declines with sequence divergence [108–117]. Importantly, there is ample empirical evidence attesting to the fact that horizontal gene transfer can affect DNA regions of various length, from fragment of genes, to multiple adjacent genes and mega-base scale extrachromosomal elements [118–126].

In sum, because of the theoretically high potential of recombination as a genetic force behind phenotypic innovation, and because of the pervasiveness of recombination in prokaryotes, I aimed to quantify the importance of recombination in metabolic innovation as one of the major goals of my dissertation.

## 1.5. Deletional robustness

It has been shown that many genes newly added to a genome via horizontal gene transfer reside in the genome only for short amounts of time [110,127]. For example, fewer than 15 percent of newly acquired genes in the *E. coli* genome may be retained in the long run [128,129]. The reason is that bacterial genomes are constantly subjected to DNA deletions on the length scales of few nucleotides to long segments of DNA comprising multiple genes [130–134]. Importantly, in bacterial genomes, deletions occur more frequently than insertions [130–135]. In other words, there is a general bias towards DNA deletion in bacterial genomes [130–135].

Moreover, recent comparative genomic studies have shown that large scale gene loss events are not a peculiarity of bacterial genomes with extremely reduced genomes such as obligate pathogens [136] or endosymbionts [137], but a dominant mode of evolution in bacterial genomes [138,139]. These observations are further confirmed by experimental evolution studies, which have shown that extensive gene loss by large-scale deletions tend to occur on short evolutionary time-scales [133,140,141]. There is evidence especially from bacterial pathogens that gene loss has not always been deleterious, but have also been associated with adaptive gains in pathogenicity

[142]. For example, the loss of allergen gene 1 (ALL1), which encodes a small cytoplasmic protein that is involved in capsule formation in *Cryptococcus neoformans* [143], of the cadA gene in *Shigella*, which encodes a Lysine decarboxylase catalyzing a reaction producing an enterotoxin inhibitor called cadaverine [144], of arabinose operon genes in *Burkholderia* [145] and of the mucA gene, encoding inhibitors of alginate biosynthetic genes in *Pseudomonas aeroginosa* [146], confer adaptive advantages during infection.

In the presence of pervasive gene loss events, in order for horizontal gene transfer to be an effective genetic force behind phenotypic innovations, bacterial genomes must ensure phenotypic robustness against the deleterious effects of such DNA deletions. Robustness is like a prerequisite for innovation; without being able to preserve current phenotypes, gaining novel phenotypes may be useless. Previous studies have shown that robustness promotes innovation in many different biological systems [147–151]. Without studying robustness, my analyses of innovation would remain incomplete. Thus, studying robustness to gene and reaction deletions in bacterial metabolic systems forms an integral part of this dissertation.

## 1.6. Bacterial genome organization

To study the robustness of bacterial genomes to large-scale gene deletions, it is necessary to study the organization of genes in bacterial genomes closely.

Comparative genomic analyses of prokaryotic and eukaryotic genomes have revealed striking differences in their genomic organization. In eukaryotes, the genome evolves predominantly through gene duplication. Chromosomes have very distinctive regions such as centromeres and telomeres and their transcription units usually include one single gene. In contrast, in prokaryotes, genomes expand mostly through horizontal gene transfer (HGT) rather than by duplication, chromosomes are relatively uniform, and genes are typically cotranscribed in operons [152].

Gene expression has strongly influenced the local organization of genes in bacterial genomes. For example, in operons enzyme-coding genes tend to be colocalized and co-transcribed in polycistronic units [153,154]. Moreover, the order of metabolic genes in the chromosome reflects the order of the corresponding enzymes in

metabolic pathways [155], and many operons code for complete metabolic pathways [156].

A fundamental question is why operons exist in prokaryotic genomes? According to the regulatory model of operon evolution, functional neighbors are adaptively brought together in the chromosome to ensure efficient gene regulation. However, this model raises two important questions. First, how do genes become closely linked before the emergence of co-regulation? Second, why are neighboring genes frequently functionally related? To address these questions, Lawrence and Roth proposed an alternative model called "the selfish operon hypothesis" [157], which claims that clustering of functionally related genes in operons is the result of selection on genes, not on organisms, to increase their own fitness through horizontal gene transfer. In other words, clustering of functionally related genes increases the probability of successful transfer, because HGT adds a complete functional module to a pre-existing network. Indeed, enzymes encoded in HGT-acquired operons have been shown to form metabolic pathways that are well integrated into the recipient metabolic network [110]. However, other studies provide evidence against the selfish operon hypothesis [158,159]. For example, operons indeed frequently contain essential genes, not just genes with peripheral metabolic functions [158]. Moreover, HGT-acquired genes have the same chance to be in operons as "native" genes, and thus there is no strong association between operons and horizontal gene transfer [159].

Above the operon level, there seem to be selective advantages in preserving the contiguity of pairs of operons [160]. Moreover, the conserved ordering of multiple operons in different genomes, which is called an uber-operon, has frequently been observed [161]. There might be regulatory advantages for contiguity between related operons. For example, operon pairs oriented in opposite directions, can share bidirectional regulatory regions, and thus allow the coregulation of the two operons [162,163]. Such operons show correlated expression levels and are more conserved than operons oriented in the same direction [164]. In addition, there might be other still unknown regulation-independent selective advantages associated with conserved operon ordering.

In addition to gene expression, DNA replication can also exert a global effect on bacterial genome organization. Because cell division is not necessarily coupled with

DNA replication, in some cases the ratio (R) between the time required to replicate the chromosome and the time between two successive cell divisions, can exceed one. Consequently cells can experience multiple replication rounds that cause a replication-induced gene dosage effect [152]. In other words, genes that are closer to the replication origin will be on average $2^R$ times more abundant in the cell than the genes close to the terminus [165]. This replication-associated gene dosage effect systematically affects the organization of genes in bacterial genomes, because highly expressed genes such as RNAP, rDNA, and the genes encoding ribosomal proteins tend to be clustered near the origin of replication [166]. Moreover, because of the asymmetric replication of DNA (i.e. the leading strand replicates continuously, whereas the lagging strand replicates semi-continuously), sequence composition and gene content on the two DNA strands can differ considerably from each other [152,167]. It has been shown that the driver of this strand bias is not gene expression, but the essentiality of genes [168,169]. Last but not least, it has also been shown that in bacterial genomes essential genes are not uniformly distributed in the genome, but they are preferentially clustered in particular regions of the genome [170,171].

In sum, in this section I described organizational principles of bacterial genomes because they can help us understand the phenotypic robustness to large-scale gene deletion that I discussed in section 1.5. As I mentioned before, studying phenotypic innovation is incomplete without studying phenotypic robustness. In the next sections I will focus on other aspects of phenotypic innovation.

## 1.7. Phenotypic constraints in biological systems

Another important aspect of phenotypic innovation is the extent of its predictability. If there is absolutely no constraint on phenotypic innovation, all possible novel phenotypes can potentially emerge, and thus phenotypic innovation becomes an unpredictable process. In contrast, if the emergence of novel phenotypes is subjected to various constraints, phenotypic innovation may be a predictable phenomenon. In this case, identifying the potential causes of phenotypic constraints can help us understand the rules governing the emergence of novel phenotypes.

As a general definition, an evolutionary constraint is a bias or limitation in the emergence of phenotypic variation in a given biological system [172]. Absolute constraints occur when some phenotypes cannot be produced, and relative constraints

exist when some phenotypes are more likely to arise than others. Are biological systems able to produce every conceivable kind of phenotypic variation? According to ample real-world examples of phenotypic constraints, it is more likely that both *absolute* and *relative* answers to this question are negative.

Ample evidence from organismal and anatomical levels down to the molecular level attests to the pervasiveness of phenotypic constraints in biological systems. Examples of absolute constraints in organismal level include the absence of photosynthesis in higher animals, the absence of birds that can give birth to live young instead of to eggs, and the absence of palm trees in cold climates [172,173]. Anatomical examples include the general lack of teeth in the lower jaw of frogs, the maximally five digits (fingers and toes) of tetrapod limbs and constrained variation in segment number, orientation and identity in the fruit fly *Drosophila melanogaster* [174]. Last but not least, molecular examples of phenotypic constraints include the absence of D-isomers in the 20 amino acids found in natural proteins [31].

There are four major causes of phenotypic constraints that are not mutually exclusive. They include physicochemical, selective, genetic, and developmental constraints. The first major case is physicochemical constraints including the limited number of protein folds caused by the packing requirements of hydrophobic amino-acids [175]. Another case in this category is the necessity to have a circulatory system in organisms above a given size to ensure efficient delivery of nutrients to all body parts [173]. The second major cause is selective constraints, which are imposed by natural selection, for example by eliminating the phenotypes with lower fitness. Third, genetic constraints occur when any one genotype and its variants can produce only a tiny fraction of all possible phenotypes. For example, while mutations in the fly *Drosophila subobscura* cause the emergence of a wide range of wing shapes and eye morphology, in *Drosophila melanogaster* mutations do not have the same broad phenotypic effects [172]. Fourth, the developmental processes that produce a phenotype from information encoded in a genotype can impose further constraints on phenotypic variation. For example, the number of digits in salamanders and frogs is constrained during development [176,177].

These anecdotal examples can inform us about the existence of phenotypic constraints in biological systems. However, in order to quantify the extent of phenotypic variation in a given biological system, we need to study thousands or

millions of genotypes with their corresponding phenotypes. The genotype-phenotype map that I establish in metabolic systems based on flux balance analysis and Markov Chain Monte Carlo sampling algorithms can thus be an invaluable resource to systematically analyze the pervasiveness of phenotypic constraints and their underlying causes.

## 1.8. Historical contingency

A concept closely related to phenotypic constrain is historical contingency. If the origin of a novel phenotype depends on the history of a population, for example, on pre-existing genotypes or phenotypes, we consider that an example of historical contingency [178,179]. Historical contingency can occur especially because of the interactions between different mutations; for example, when the beneficial effect of a mutation does not manifest itself unless it occurs with another mutation, we can speak of contingent mutation effects [180,181]. As a prominent example of historical contingency, experimental evolution of *Escherichia coli* in Lenski lab has shown that the emergence of *E. coli* strains with the ability to utilize citrate as a novel metabolic phenotype is strongly dependent on the occurrence of potentiating mutations after 20,000 generations [182]. In other words, this long-term evolutionary experiment has shown that the emergence of novel citrate utilization phenotypes is strongly contingent on the genetic history of a population [182].

In the framework of genotype networks, historical contingency can be systematically quantified. Of special importance in this context is the concept of connectivity in genotype networks. If a genotype network, which is comprised of genotypes with the same phenotype, forms a single connected component in genotype space, where every pair of genotypes is reachable from each other through a phenotype-preserving mutational path, historical contingency will not play a major role in the phenotypic outcome of genotypic changes in the system. In contrast, if a genotype network is fragmented into several independent connected components, the evolutionary fate of a system may strongly depend on the initial genotypes and the mutational history of the system. Moreover, fragmentation of genotype space can potentially restrict the accessibility of novel phenotypes.

Previous studies on genotype networks in RNA molecules have shown that RNA genotype networks are highly fragmented, therefore historical contingency can play a

major role in the evolution of RNA secondary structure phenotypes [183,184]. The situation is even more extreme in gene regulatory networks, where genotype networks can be fragmented into $10^8$ distinct connected components [185]. However, the role of historical contingency in metabolic systems is not well studied. Understanding its extent was therefore one of my research goals.

## 1.9. Exaptation

There are two broad candidates for the origin of novel phenotypes or traits. First, such traits can originate as adaptations that help an organism survive or reproduce. Second, they can also have non-adaptive origins as pre-adaptations or exaptations [186,187]. Exaptation occurs when a trait, which served initially an old function, evolves to serve a new function. In other words, the function of a trait shifts during its evolutionary history. The original idea of exaptation can be attributed to Darwin, where in *Origin* he said that "an organ originally constructed for one purpose… may be converted to one for a widely different purpose" [2].

Multiple lines of evidence from the organismal down to the molecular scale later confirmed the importance of exaptations as potential sources of evolutionary innovations [188–190]. The classical example of exaptation is feathers, which are made of keratins and originally served as thermoregulation and waterproofing, but were later "exapted" for flight [186]. Another anatomical example is lungs in ancient extant fishes that underwent exaptation to become gas bladder in present-day fishes [191]. Exaptation may have also played an important role in human evolution [192]. Molecular examples of exaptation include crystallins, which originally were metabolic enzymes and later became light-refracting proteins in eye lenses [193]. Gene regulation is another potential source of exaptation. The same gene or set of genes can serve different functions by changing their patterns of regulations [188]. Moreover, metabolic networks, by possessing an inherent flexibility in utilizing a wide variety of substrates as sources of energy, show high capacity for exaptation to new environments [194].

Despite the widespread examples of exaptations in life, it is not clear to what extent exaptation can contribute to evolutionary innovations. In other words, a systematic quantification of the potential of a biological system for exaptation is required. Large-scale quantitative genotype-phenotype maps in metabolic systems can provide us

with an unprecedented opportunity to approach this problem. A previous study quantified exaptation in genome-scale metabolic networks based on sampling methods [194]. As the last objective of my dissertation, here, I complemented that study using the exhaustive genotype-phenotype map of central carbon metabolism.

## 1.10. Thesis outline

I present the results of my research in 6 chapters. The first one is **chapter 2**, where I describe my quantitative analyses on the role of recombination for phenotypic innovation in genome-scale metabolic networks. My observations reveal the power of recombination in originating metabolic innovations. I systematically characterize genotypic and phenotypic features of recombining parental metabolisms that can enhance the emergence of innovative offspring. Moreover, my results highlight the importance of a specific class of reactions called super-essential reactions in metabolic innovations.

In **chapter 3**, I focus on the phenotypic robustness that is a prerequisite for phenotypic innovation. I show that bacterial genomes have evolved a genomic organization that provides a substantially higher robustness to large-scale metabolic gene deletions. I follow these observations with systematic analyses of the evolutionary forces that can create such a genomic organization. I show that essential genes are significantly clustered in bacterial genome, and I provide empirical evidence implying that this gene clustering might be an adaptive response to the pervasive gene loss events that bacterial genomes are exposed to.

In **chapter 4**, using recombining parental metabolic networks with specific phenotypes, I comprehensively quantify the extent of phenotypic constraints and contingencies in complex metabolic systems. My results reveal that the emergence of novel phenotypes in metabolic systems is not absolutely but only relatively constrained by and contingent on parental phenotypes or genotypes. Moreover, I suggest biochemical causes behind such phenotypic constraints and contingencies.

In **chapter 5, 6 and 7**, I exclusively focus on central carbon metabolism and construct an exhaustive genotype-phenotype map comprising $10^{15}$ genotypic variants. In **chapter 5**, by analyzing the connectivity of genotype networks, I show that

historical contingency does not play a substantial role in the evolution of metabolic properties in central carbon metabolism. In **chapter 6**, I describe how the organization of genotypes in the genotype space facilitates the emergence of novel phenotypes in central carbon metabolism. Finally, the analyses of **chapter 7** reveal a high potential for exaptation as a non-adaptive origin of evolutionary innovations in central carbon metabolism.

## 1.11. References


1. Wagner A (2014) Arrival of the Fittest: Solving Evolution's Greatest Puzzle. 1st ed. London: Oneworld Publications.
2. Darwin C (1872) Darwin Online: On the Origin of Species. 6th ed. Murray M, editor London: Adamant Media Corporation. Available: http://darwin-online.org.uk/EditorialIntroductions/Freeman_OntheOriginofSpecies.html.
3. Wagner A (2011) The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems. Oxford: Oxford University Press. 132-143 p. Available: http://www.amazon.com/The-Origins-Evolutionary-Innovations-Transformative/dp/0199692602. Accessed 13 September 2015.
4. Rehmann L, Daugulis AJ (2008) Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. Biotechnol Bioeng 99: 521–528. doi:10.1002/bit.21610.
5. van der Meer JR, Werlen C, Nishino S, Spain J (1998) Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. Appl Environ Microbiol 64: 4185–4193. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=106626&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.
6. Cline RE, Hill RH, Phillips DL, Needham LL (n.d.) Pentachlorophenol measurements in body fluids of people in log homes and workplaces. Arch Environ Contam Toxicol 18: 475–481.
7. Copley SD (2000) Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. Trends Biochem Sci 25: 261–265.
8. Dantas G, Sommer MOA, Oluwasegun RD, Church GM (2008) Bacteria subsisting on antibiotics. Science 320: 100–103. doi:10.1126/science.1155157.
9. Detkova EN, Boltyanskaya Y V. (2007) Osmoadaptation of haloalkaliphilic bacteria: Role of osmoregulators and their possible practical application. Microbiology 76: 511–522. doi:10.1134/S0026261707050013.
10. Boltyanskaya Y V., Detkova EN, Shumskii AN, Dulov LE, Pusheva MA (2005) Osmoadaptation in Representatives of Haloalkaliphilic Bacteria from Soda Lakes. Microbiology 74: 640–645. Available: http://link.springer.com/10.1007/s11021-005-0117-5. Accessed 15 October 2017.
11. Detkova EN, Pusheva MA (2006) Energy metabolism in halophilic and

alkaliphilic acetogenic bacteria. Microbiology 75: 1–11. Available: http://link.springer.com/10.1134/S0026261706010012. Accessed 15 October 2017.

12. Rothschild LJ (2008) The evolution of photosynthesis…again? Philos Trans R Soc London B Biol Sci 363. Available: http://rstb.royalsocietypublishing.org/content/363/1504/2787.long. Accessed 17 October 2017.

13. Xiong J, Bauer CE (2002) Complex evolution of photosynthesis. Annu Rev Plant Biol 53: 503–521. Available: http://www.ncbi.nlm.nih.gov/pubmed/12221987. Accessed 17 October 2017.

14. Takiguchi M, Matsubasa T, Amaya Y, Mori M (1989) Evolutionary aspects of urea cycle enzyme genes. Bioessays 10: 163–166. Available: http://www.ncbi.nlm.nih.gov/pubmed/2662961. Accessed 17 February 2014.

15. Pichersky E, Gang DR (2000) Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. Trends Plant Sci 5: 439–445. Available: http://www.ncbi.nlm.nih.gov/pubmed/11044721. Accessed 2 November 2017.

16. Dudareva N, Pichersky E (2000) Biochemical and Molecular Genetic Aspects of Floral Scents. Plant Physiol 122. Available: http://www.plantphysiol.org/content/122/3/627. Accessed 2 November 2017.

17. How genes paint flowers and seeds (1998). Trends Plant Sci 3: 212–217. Available: http://www.sciencedirect.com/science/article/pii/S1360138598012424. Accessed 2 November 2017.

18. Bennet RN, Wallsgrove RM (1994) Secondary metabolites in plant defence mechanisms. New Phytol 127: 617–633. Available: http://doi.wiley.com/10.1111/j.1469-8137.1994.tb02968.x. Accessed 2 November 2017.

19. The comparative biochemistry of phytoalexin induction in plants (1999). Biochem Syst Ecol 27: 335–367. Available: http://www.sciencedirect.com/science/article/pii/S0305197898000957. Accessed 2 November 2017.

20. Trossat C, Rathinasabapathi B, Weretilnyk EA, Shen TL, Huang ZH, et al. (1998) Salinity promotes accumulation of 3-dimethylsulfoniopropionate and its precursor S-methylmethionine in chloroplasts. Plant Physiol 116: 165–171. Available: http://www.ncbi.nlm.nih.gov/pubmed/9449841. Accessed 2 November 2017.

21. Nuccio ML, Rhodes D, McNeil SD, Hanson AD (1999) Metabolic engineering of plants for osmotic stress resistance. Curr Opin Plant Biol 2: 128–134. Available: http://www.ncbi.nlm.nih.gov/pubmed/10322193. Accessed 2 November 2017.

22. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol Syst Biol 9: 661. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3658273&tool=pm

centrez&rendertype=abstract. Accessed 27 March 2015.

23. Orth JD, Fleming RMT, Palsson BØ (2010) Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal Plus 1. doi:10.1128/ecosalplus.10.2.1.

24. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409&tool=pmcentrez&rendertype=abstract. Accessed 10 July 2014.

25. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-60. doi:10.1093/nar/gkp896.

26. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354-7. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1347464&tool=pmcentrez&rendertype=abstract. Accessed 6 August 2015.

27. King ZA, Lu J, Dräger A, Miller P, Federowicz S, et al. (2015) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res: gkv1049-. Available: http://nar.oxfordjournals.org/content/early/2015/10/15/nar.gkv1049. Accessed 18 October 2015.

28. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28: 977–982. Available: http://dx.doi.org/10.1038/nbt.1672. Accessed 9 July 2014.

29. Wagner A (2012) Metabolic networks and their evolution. Adv Exp Med Biol 751: 29–52. doi:10.1007/978-1-4614-3567-9_2.

30. Zielinski DC, Palsson BØ (2012) Kinetic Modeling of Metabolic Networks. Systems Metabolic Engineering. Dordrecht: Springer Netherlands. pp. 25–55. Available: http://www.springerlink.com/index/10.1007/978-94-007-4534-6_2. Accessed 18 October 2017.

31. Nelson DL, Cox MM (2004) Lehninger Principles of Biochemistry. 3rd ed. New York: W. H. Freeman. Available: https://www.amazon.com/dp/1464126119/ref=rdr_ext_tmb.

32. Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19: 125–130. doi:10.1038/84379.

33. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. Curr Opin Biotechnol 14: 491–496. Available: http://www.ncbi.nlm.nih.gov/pubmed/14580578. Accessed 21 September 2014.

34. Schuster S, Dandekar T, Fell DA (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends Biotechnol 17: 53–60. Available: http://www.ncbi.nlm.nih.gov/pubmed/10087604. Accessed 18 November

2017.

35. Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C (2013) Elementary flux modes in a nutshell: Properties, calculation and applications. Biotechnol J 8: 1009–1016. Available: http://www.ncbi.nlm.nih.gov/pubmed/23788432. Accessed 18 November 2017.

36. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat Biotechnol 18: 326–332. Available: http://www.ncbi.nlm.nih.gov/pubmed/10700151. Accessed 18 November 2017.

37. Trinh CT, Wlaschin A, Srienc F (2009) Elementary mode analysis: a useful metabolic pathway analysis tool for characterizing cellular metabolism. Appl Microbiol Biotechnol 81: 813–826. Available: http://www.ncbi.nlm.nih.gov/pubmed/19015845. Accessed 19 November 2017.

38. Carlson R, Srienc F (2004) FundamentalEscherichia coli biochemical pathways for biomass and energy production: creation of overall flux states. Biotechnol Bioeng 86: 149–162. Available: http://www.ncbi.nlm.nih.gov/pubmed/15052634. Accessed 19 November 2017.

39. Carlson R, Srienc F (2004) FundamentalEscherichia coli biochemical pathways for biomass and energy production: Identification of reactions. Biotechnol Bioeng 85: 1–19. Available: http://www.ncbi.nlm.nih.gov/pubmed/14705007. Accessed 19 November 2017.

40. Trinh CT, Carlson R, Wlaschin A, Srienc F (2006) Design, construction and performance of the most efficient biomass producing *E. coli* bacterium. Metab Eng 8: 628–638. Available: http://www.ncbi.nlm.nih.gov/pubmed/16997589. Accessed 19 November 2017.

41. Unrean P, Trinh CT, Srienc F (2010) Rational design and construction of an efficient *E. coli* for production of diapolycopendioic acid. Metab Eng 12: 112–122. Available: http://www.ncbi.nlm.nih.gov/pubmed/19944775. Accessed 19 November 2017.

42. Trinh CT, Srienc F (2009) Metabolic Engineering of Escherichia coli for Efficient Conversion of Glycerol to Ethanol. Appl Environ Microbiol 75: 6696–6705. Available: http://www.ncbi.nlm.nih.gov/pubmed/19734340. Accessed 19 November 2017.

43. Trinh CT, Unrean P, Srienc F (2008) Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. Appl Environ Microbiol 74: 3634–3643. Available: http://www.ncbi.nlm.nih.gov/pubmed/18424547. Accessed 19 November 2017.

44. Trinh CT, Li J, Blanch HW, Clark DS (2011) Redesigning Escherichia coli metabolism for anaerobic production of isobutanol. Appl Environ Microbiol 77: 4894–4904. Available: http://www.ncbi.nlm.nih.gov/pubmed/21642415.

Accessed 19 November 2017.

45. Terzer M, Stelling J (2008) Large-scale computation of elementary flux modes with bit pattern trees. Bioinformatics 24: 2229–2235. Available: http://www.ncbi.nlm.nih.gov/pubmed/18676417. Accessed 19 November 2017.

46. Kamp A v., Schuster S (2006) Metatool 5.0: fast and flexible elementary modes analysis. Bioinformatics 22: 1930–1931. Available: http://www.ncbi.nlm.nih.gov/pubmed/16731697. Accessed 19 November 2017.

47. Rocha I, Maia P, Evangelista P, Vilaça P, Soares S, et al. (2010) OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol 4: 45. Available: http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-4-45. Accessed 19 November 2017.

48. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, et al. (2006) COPASI--a COmplex PAthway SImulator. Bioinformatics 22: 3067–3074. Available: http://www.ncbi.nlm.nih.gov/pubmed/17032683. Accessed 19 November 2017.

49. Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED (2006) A methodology for the structural and functional analysis of signaling and regulatory networks. BMC Bioinformatics 7: 56. Available: http://www.ncbi.nlm.nih.gov/pubmed/16464248. Accessed 19 November 2017.

50. Schwarz R, Musch P, von Kamp A, Engels B, Schirmer H, et al. (2005) YANA – a software tool for analyzing flux modes, gene-expression and enzyme activities. BMC Bioinformatics 6: 135. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-135. Accessed 19 November 2017.

51. Gagneur J, Klamt S (2004) Computation of elementary modes: a unifying framework and the new binary approach. BMC Bioinformatics 5: 175. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-175. Accessed 19 November 2017.

52. Klamt S, Gagneur J, von Kamp A (2005) Algorithmic approaches for computing elementary modes in large biochemical reaction networks. Syst Biol (Stevenage) 152: 249–255. Available: http://www.ncbi.nlm.nih.gov/pubmed/16986267. Accessed 19 November 2017.

53. Terzer M, Stelling J (2006) Accelerating the Computation of Elementary Modes Using Pattern Trees Springer, Berlin, Heidelberg. pp. 333–343. Available: http://link.springer.com/10.1007/11851561_31. Accessed 19 November 2017.

54. Jevremović D, Trinh CT, Srienc F, Sosa CP, Boley D (2011) Parallelization of Nullspace Algorithm for the computation of metabolic pathways. Parallel Comput 37: 261–278. Available: http://www.ncbi.nlm.nih.gov/pubmed/22058581. Accessed 19 November 2017.

55. Kaleta C, Filipe De Figueiredo L, Behre J, Schuster S (n.d.) EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.466.3553&rep=rep1&type=pdf. Accessed 19 November 2017.

56. Machado D, Soons Z, Patil KR, Ferreira EC, Rocha I (2012) Random sampling of elementary flux modes in large-scale metabolic networks. Bioinformatics 28: i515–i521. Available: http://www.ncbi.nlm.nih.gov/pubmed/22962475. Accessed 19 November 2017.

57. Tang YJ, Martin HG, Myers S, Rodriguez S, Baidoo EEK, et al. (2009) Advances in analysis of microbial metabolic fluxes via $^{13}$C isotopic labeling. Mass Spectrom Rev 28: 362–375. Available: http://www.ncbi.nlm.nih.gov/pubmed/19025966. Accessed 21 October 2017.

58. Calheiros Gomes L, Simoes M (2012) 13C Metabolic Flux Analysis: From the Principle to Recent Applications. Curr Bioinform 7: 77–86. Available: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1574-8936&volume=7&issue=1&spage=77. Accessed 21 October 2017.

59. Sauer U (2006) Metabolic networks in motion: 13 C-based flux analysis. Mol Syst Biol. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1682028/pdf/msb4100109.pdf. Accessed 21 October 2017.

60. Sauer U (2004) High-throughput phenomics: experimental methods for mapping fluxomes. Curr Opin Biotechnol 15: 58–63. Available: http://www.ncbi.nlm.nih.gov/pubmed/15102468. Accessed 21 October 2017.

61. Wiechert W (2001) 13C Metabolic Flux Analysis. Metab Eng 3: 195–206. Available: http://www.ncbi.nlm.nih.gov/pubmed/11461141. Accessed 21 October 2017.

62. Marx A, de Graaf AA, Wiechert W, Eggeling L, Sahm H (1996) Determination of the fluxes in the central metabolism ofCorynebacterium glutamicum by nuclear magnetic resonance spectroscopy combined with metabolite balancing. Biotechnol Bioeng 49: 111–129. Available: http://www.ncbi.nlm.nih.gov/pubmed/18623562. Accessed 21 October 2017.

63. Sauer U, Hatzimanikatis V, Bailey JE, Hochuli M, Szyperski T, et al. (1997) Metabolic fluxes in riboflavin-producing Bacillus subtilis. Nat Biotechnol 15: 448–452. Available: http://www.ncbi.nlm.nih.gov/pubmed/9131624. Accessed 21 October 2017.

64. Portais J-C, Delort A-M (2002) Carbohydrate cycling in micro-organisms: what can (13)C-NMR tell us? FEMS Microbiol Rev 26: 375–402. Available: http://www.ncbi.nlm.nih.gov/pubmed/12413666. Accessed 21 October 2017.

65. Gombert AK, Moreira dos Santos M, Christensen B, Nielsen J (2001) Network Identification and Flux Quantification in the Central Metabolism of Saccharomyces cerevisiae under Different Conditions of Glucose Repression. J Bacteriol 183: 1441–1451. Available: http://www.ncbi.nlm.nih.gov/pubmed/11157958. Accessed 21 October 2017.

66. Fischer E, Sauer U (2003) Metabolic flux profiling of Escherichia coli mutants in central carbon metabolism using GC-MS. Eur J Biochem 270: 880–891. Available: http://doi.wiley.com/10.1046/j.1432-1033.2003.03448.x. Accessed 21 October 2017.

67. Klapa MI, Aon J-C, Stephanopoulos G (2003) Systematic quantification of complex metabolic flux networks using stable isotopes and mass spectrometry. Eur J Biochem 270: 3525–3542. Available: http://www.ncbi.nlm.nih.gov/pubmed/12919317. Accessed 21 October 2017.

68. Szyperski T (1995) Biosynthetically directed fractional 13C-labeling of proteinogenic amino acids. An efficient analytical tool to investigate intermediary metabolism. Eur J Biochem 232: 433–448. Available: http://www.ncbi.nlm.nih.gov/pubmed/7556192. Accessed 21 October 2017.

69. Becker J, Klopprogge C, Herold A, Zelder O, Bolten CJ, et al. (2007) Metabolic flux engineering of l-lysine production in Corynebacterium glutamicum—over expression and modification of G6P dehydrogenase. J Biotechnol 132: 99–109. Available: http://www.ncbi.nlm.nih.gov/pubmed/17624457. Accessed 21 October 2017.

70. Shirai T, Fujimura K, Furusawa C, Nagahisa K, Shioya S, et al. (2007) Study on roles of anaplerotic pathways in glutamate overproduction of Corynebacterium glutamicum by metabolic flux analysis. Microb Cell Fact 6: 19. Available: http://www.ncbi.nlm.nih.gov/pubmed/17587457. Accessed 21 October 2017.

71. Kleijn RJ, Liu F, van Winden WA, van Gulik WM, Ras C, et al. (2007) Cytosolic NADPH metabolism in penicillin-G producing and non-producing chemostat cultures of Penicillium chrysogenum. Metab Eng 9: 112–123. Available: http://www.ncbi.nlm.nih.gov/pubmed/17008114. Accessed 21 October 2017.

72. Fell D. (1997) Understanding the control of metabolism. Portland Press. 301 p.

73. Kacser H, Burns JA (1973) The control of flux. Symp Soc Exp Biol 27: 65–104. Available: http://www.ncbi.nlm.nih.gov/pubmed/4148886. Accessed 22 October 2017.

74. Heinrich R, Rapoport TA (1974) A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. Eur J Biochem 42: 89–95. Available: http://www.ncbi.nlm.nih.gov/pubmed/4830198. Accessed 22 October 2017.

75. Westerhoff H V, Chen YD (1984) How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control. Eur J Biochem 142: 425–430. Available: http://www.ncbi.nlm.nih.gov/pubmed/6745283. Accessed 22 October 2017.

76. Wildermuth MC (2000) Metabolic control analysis: biological applications and insights. Genome Biol 1: REVIEWS1031. Available: http://www.ncbi.nlm.nih.gov/pubmed/11178271. Accessed 22 October 2017.

77. Moreno-Sánchez R, Saavedra E, Rodríguez-Enríquez S, Olín-Sandoval V (2008) Metabolic control analysis: a tool for designing strategies to manipulate metabolic pathways. J Biomed Biotechnol 2008: 597913. Available:

http://www.ncbi.nlm.nih.gov/pubmed/18629230. Accessed 22 October 2017.

78. Cascante M, Boros LG, Comin-Anduix B, de Atauri P, Centelles JJ, et al. (2002) Metabolic control analysis in drug discovery and disease. Nat Biotechnol 20: 243–249. Available: http://www.nature.com/doifinder/10.1038/nbt0302-243. Accessed 22 October 2017.

79. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186–189. doi:10.1038/nature01149.

80. Fong SS, Marciniak JY, Palsson BO (2003) Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J Bacteriol 185: 6400–6408. doi:10.1128/JB.185.21.6400-6408.2003.

81. Fong SS, Palsson BØ (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36: 1056–1058. doi:10.1038/ng1432.

82. Noor E, Eden E, Milo R, Alon U (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell 39: 809–820. doi:10.1016/j.molcel.2010.08.031.

83. Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci 255: 279–284. Available: http://www.ncbi.nlm.nih.gov/pubmed/7517565. Accessed 3 January 2016.

84. Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, et al. (1996) Analysis of RNA sequence structure maps by exhaustive enumeration .1. Neutral networks. Monatshefte für Chemie 127: 374. Available: http://pub.uni-bielefeld.de/publication/1639477. Accessed 20 September 2014.

85. Newman M (2010) Networks: An Introduction. Oxford: Oxford University Press. 784 p. Available: http://books.google.com/books?id=LrFaU4XCsUoC&pgis=1. Accessed 21 September 2014.

86. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of State Calculations by Fast Computing Machines. J Chem Phys 21: 1087–1092. Available: http://aip.scitation.org/doi/10.1063/1.1699114. Accessed 22 October 2017.

87. Samal A, Matias Rodrigues JF, Jost J, Martin OC, Wagner A (2010) Genotype networks in metabolic reaction spaces. BMC Syst Biol 4: 30. doi:10.1186/1752-0509-4-30.

88. Matias Rodrigues JF, Wagner A (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol 5: e1000613. Available: http://dx.plos.org/10.1371/journal.pcbi.1000613. Accessed 17 February 2014.

89. Shapir N, Mongodin EF, Sadowsky MJ, Daugherty SC, Nelson KE, et al. (2007) Evolution of catabolic pathways: Genomic insights into microbial s-triazine metabolism. J Bacteriol 189: 674–682. Available:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1797303&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

90. Cohen J (2001) How DNA Shuffling Works. Science (80- ) 293. Available: http://science.sciencemag.org/content/293/5528/237. Accessed 22 October 2017.

91. Stemmer WP (1994) DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc Natl Acad Sci 91: 10747–10751. Available: http://www.pnas.org/content/91/22/10747. Accessed 16 July 2015.

92. Crameri A, Raillard SA, Bermudez E, Stemmer WP (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature 391: 288–291. Available: http://dx.doi.org/10.1038/34663. Accessed 13 September 2015.

93. Zhang Y-X, Perry K, Vinci VA, Powell K, Stemmer WPC, et al. (2002) Genome shuffling leads to rapid phenotypic improvement in bacteria. Nature 415: 644–646. Available: http://dx.doi.org/10.1038/415644a. Accessed 13 September 2015.

94. Allen EE, Tyson GW, Whitaker RJ, Detter JC, Richardson PM, et al. (2007) Genome dynamics in a natural archaeal population. Proc Natl Acad Sci U S A 104: 1883–1888. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1794283&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

95. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, et al. (2001) Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A 98: 182–187. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=14565&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

96. Guttman DS, Dykhuizen DE (1994) Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science 266: 1380–1383. Available: http://www.ncbi.nlm.nih.gov/pubmed/7973728. Accessed 17 February 2014.

97. Whitaker RJ, Grogan DW, Taylor JW (2005) Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol Biol Evol 22: 2354–2361. Available: http://www.ncbi.nlm.nih.gov/pubmed/16093568. Accessed 28 January 2014.

98. Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF (2004) Frequent recombination in a saltern population of Halorubrum. Science 306: 1928–1929. Available: http://www.ncbi.nlm.nih.gov/pubmed/15591201. Accessed 29 January 2014.

99. Majewski J (2001) Sexual isolation in bacteria. FEMS Microbiol Lett 199: 161–169. Available: http://www.ncbi.nlm.nih.gov/pubmed/11377861. Accessed 17 February 2014.

100. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315: 476–480. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2220085&tool=pm

centrez&rendertype=abstract. Accessed 26 January 2014.

101. Coleman ML, Chisholm SW (2010) Ecosystem-specific selection pressures revealed through comparative population genomics. Proc Natl Acad Sci U S A 107: 18634–18639. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2972931&tool=pmcentrez&rendertype=abstract. Accessed 23 January 2014.

102. Denef VJ, Kalnejais LH, Mueller RS, Wilmes P, Baker BJ, et al. (2010) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proc Natl Acad Sci U S A 107: 2383–2390. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2823883&tool=pmcentrez&rendertype=abstract. Accessed 23 January 2014.

103. Eppley JM, Tyson GW, Getz WM, Banfield JF (2007) Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics 177: 407–416. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2013692&tool=pmcentrez&rendertype=abstract. Accessed 14 February 2014.

104. Lo I, Denef VJ, Verberkmoes NC, Shah MB, Goltsman D, et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. Nature 446: 537–541. Available: http://www.ncbi.nlm.nih.gov/pubmed/17344860. Accessed 23 January 2014.

105. Nesbø CL, Dlutek M, Doolittle WF (2006) Recombination in Thermotoga: implications for species concepts and biogeography. Genetics 172: 759–769. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1456242&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

106. Denef VJ, Banfield JF (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. Science 336: 462–466. Available: http://www.ncbi.nlm.nih.gov/pubmed/22539719. Accessed 7 July 2016.

107. Koehler A, Karch H, Beikler T, Flemmig TF, Suerbaum S, et al. (2003) Multilocus sequence analysis of Porphyromonas gingivalis indicates frequent recombination. Microbiology 149: 2407–2415. Available: http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.26267-0. Accessed 7 July 2016.

108. Thomas CM, Nielsen KM (2005) Mechanisms of, and Barriers to, Horizontal Gene Transfer between Bacteria. Nat Rev Microbiol 3: 711–721. Available: http://www.ncbi.nlm.nih.gov/pubmed/16138099. Accessed 17 July 2017.

109. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399: 323–329. Available: http://www.ncbi.nlm.nih.gov/pubmed/10360571. Accessed 17 February 2014.

110. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375. doi:10.1038/ng1686.

111. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary Origins of Genomic Repertoires in Bacteria. PLoS Biol 3: e130. doi:10.1371/journal.pbio.0030130.

112. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304. Available: http://www.ncbi.nlm.nih.gov/pubmed/10830951. Accessed 22 January 2014.

113. Choi I-G, Kim S-H (2007) Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A 104: 4489–4494. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1815472&tool=pmcentrez&rendertype=abstract. Accessed 10 February 2014.

114. Daubin V, Ochman H (2004) Quartet mapping and the extent of lateral transfer in bacterial genomes. Mol Biol Evol 21: 86–89. Available: http://www.ncbi.nlm.nih.gov/pubmed/12949130. Accessed 10 February 2014.

115. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, et al. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 464: 908–912. Available: http://www.ncbi.nlm.nih.gov/pubmed/20376150. Accessed 21 January 2014.

116. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, et al. (2011) Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480: 241–244. Available: http://www.ncbi.nlm.nih.gov/pubmed/22037308. Accessed 22 January 2014.

117. Wagner A, de la Chaux N (2008) Distant horizontal gene transfer is rare for multiple families of prokaryotic insertion sequences. Mol Genet Genomics 280: 397–408. Available: http://www.ncbi.nlm.nih.gov/pubmed/18751731. Accessed 17 February 2014.

118. Lin CH, Bourque G, Tan P (2008) A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. Mol Biol Evol 25: 549–558. Available: http://www.ncbi.nlm.nih.gov/pubmed/18162473. Accessed 24 June 2016.

119. Bork P, Doolittle RF (1992) Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci 89: 8990–8994. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.89.19.8990. Accessed 24 June 2016.

120. Inagaki Y, Susko E, Roger AJ (2006) Recombination between elongation factor 1 genes from distantly related archaeal lineages. Proc Natl Acad Sci 103: 4528–4533. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0600744103. Accessed 24 June 2016.

121. Hartl DL, Lozovskaya ER, Lawrence JG (1992) Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica 86: 47–53. Available: http://www.ncbi.nlm.nih.gov/pubmed/1334917. Accessed 24 June 2016.

122. Igarashi N, Harada J, Nagashima S, Matsuura K, Shimada K, et al. (2001) Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J Mol Evol 52: 333–341. Available: http://www.ncbi.nlm.nih.gov/pubmed/11343129. Accessed 24 June 2016.

123. Omelchenko M V, Makarova KS, Wolf YI, Rogozin IB, Koonin E V (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol 4: R55. Available: http://www.ncbi.nlm.nih.gov/pubmed/12952534. Accessed 24 June 2016.

124. Chan CX, Beiko RG, Darling AE, Ragan MA (2010) Lateral Transfer of Genes and Gene Fragments in Prokaryotes. Genome Biol Evol 1: 429–438. Available: http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evp044. Accessed 24 June 2016.

125. Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, et al. (2000) Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell 103: 711–721. Available: http://www.ncbi.nlm.nih.gov/pubmed/11114328. Accessed 24 June 2016.

126. Didelot X, Achtman M, Parkhill J, Thomson NR, Falush D (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res 17: 61–68. Available: http://www.ncbi.nlm.nih.gov/pubmed/17090663. Accessed 24 June 2016.

127. Lawrence JG, Ochman H (1998) Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A 95: 9413–9417. Available: http://www.ncbi.nlm.nih.gov/pubmed/9689094. Accessed 22 October 2017.

128. Ochman H, Davalos LM (2006) The Nature and Dynamics of Bacterial Genomes. Science (80- ) 311: 1730–1733. Available: http://www.ncbi.nlm.nih.gov/pubmed/16556833. Accessed 22 October 2017.

129. Kuo C-H, Ochman H (2009) The fate of new bacterial genes. FEMS Microbiol Rev 33: 38–43. Available: http://www.ncbi.nlm.nih.gov/pubmed/19054121. Accessed 30 January 2014.

130. Andersson JO, Andersson SG (2001) Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol Biol Evol 18: 829–839. Available: http://www.ncbi.nlm.nih.gov/pubmed/11319266. Accessed 17 February 2014.

131. Kuo C-H, Ochman H (2009) Deletional bias across the three domains of life. Genome Biol Evol 1: 145–152. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2817411&tool=pmcentrez&rendertype=abstract. Accessed 22 January 2014.

132. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17: 589–596. Available: http://www.ncbi.nlm.nih.gov/pubmed/11585665. Accessed 4 February 2014.

133. Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD, et al. (2005) Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci U S A 102: 12112–12116. Available: http://www.ncbi.nlm.nih.gov/pubmed/16099836. Accessed 17 July 2017.

134. van Passel MWJ, Marri PR, Ochman H (2008) The emergence and fate of horizontally acquired genes in Escherichia coli. PLoS Comput Biol 4: e1000059. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2275313&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

135. Kunin V, Ouzounis CA (2003) The Balance of Driving Forces During Genome Evolution in Prokaryotes. Genome Res 13: 1589–1594. Available: http://www.ncbi.nlm.nih.gov/pubmed/12840037. Accessed 17 July 2017.

136. McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10: 13. Available: http://www.nature.com/doifinder/10.1038/nrmicro2670. Accessed 17 July 2017.

137. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. Cell 108: 583–586. Available: http://www.ncbi.nlm.nih.gov/pubmed/11893328. Accessed 17 July 2017.

138. Wolf YI, Koonin E V. (2013) Genome reduction as the dominant mode of evolution. BioEssays 35: 829–837. Available: http://www.ncbi.nlm.nih.gov/pubmed/23801028. Accessed 17 July 2017.

139. Albalat R, Cañestro C (2016) Evolution by gene loss. Nat Rev Genet 17: 379–391. Available: http://www.ncbi.nlm.nih.gov/pubmed/27087500. Accessed 17 July 2017.

140. Lee M-C, Marx CJ, Lenski R, Sivam D, Lidstrom M (2012) Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. PLoS Genet 8: e1002651. Available: http://dx.plos.org/10.1371/journal.pgen.1002651. Accessed 17 July 2017.

141. Koskiniemi S, Sun S, Berg OG, Andersson DI, Boxer D (2012) Selection-Driven Gene Loss in Bacteria. PLoS Genet 8: e1002787. Available: http://dx.plos.org/10.1371/journal.pgen.1002787. Accessed 17 July 2017.

142. Sokurenko E V, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. Trends Microbiol 7: 191–195. Available: http://www.ncbi.nlm.nih.gov/pubmed/10354593. Accessed 17 July 2017.

143. Jain N, Li L, Hsueh Y-P, Guerrero A, Heitman J, et al. (2009) Loss of allergen 1 confers a hypervirulent phenotype that resembles mucoid switch variants of Cryptococcus neoformans. Infect Immun 77: 128–140. Available: http://www.ncbi.nlm.nih.gov/pubmed/18955480. Accessed 22 October 2017.

144. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. Proc Natl Acad Sci U S A 95: 3943–3948. Available: http://www.ncbi.nlm.nih.gov/pubmed/9520472. Accessed 17 July 2017.

145. Moore RA, Reckseidler-Zenteno S, Kim H, Nierman W, Yu Y, et al. (2004) Contribution of Gene Loss to the Pathogenic Evolution of Burkholderia pseudomallei and Burkholderia mallei. Infect Immun 72: 4172–4187. Available: http://www.ncbi.nlm.nih.gov/pubmed/15213162. Accessed 17 July 2017.

146. Yu H, Hanes M, Chrisp CE, Boucher JC, Deretic V (1998) Microbial pathogenesis in cystic fibrosis: pulmonary clearance of mucoid Pseudomonas aeruginosa and inflammation in a mouse model of repeated respiratory challenge. Infect Immun 66: 280–288. Available:

http://www.ncbi.nlm.nih.gov/pubmed/9423869. Accessed 22 October 2017.

147. Fares MA (2015) Survival and innovation: The role of mutational robustness in evolution. Biochimie 119: 254–261. Available: http://www.ncbi.nlm.nih.gov/pubmed/25447135. Accessed 29 December 2015.

148. Wagner A (2012) The role of robustness in phenotypic adaptation and innovation. Proc R Soc B Biol Sci 279: 1249–1258. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3282381&tool=pmcentrez&rendertype=abstract. Accessed 8 October 2015.

149. Ferrada E, Wagner A (2008) Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proc R Soc London B Biol Sci 275. Available: http://rspb.royalsocietypublishing.org/content/275/1643/1595. Accessed 22 October 2017.

150. Wagner A (2008) Gene duplications, robustness and evolutionary innovations. BioEssays 30: 367–373. Available: http://www.ncbi.nlm.nih.gov/pubmed/18348184. Accessed 22 October 2017.

151. Ciliberti S, Martin OC, Wagner A (2007) Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci 104: 13591–13596. Available: http://www.pnas.org/cgi/doi/10.1073/pnas.0705396104. Accessed 11 November 2016.

152. Rocha EPC (2008) The Organization of the Bacterial Genome. Annu Rev Genet 42: 211–233. Available: http://www.ncbi.nlm.nih.gov/pubmed/18605898. Accessed 23 October 2017.

153. Jacob F, Perrin D, Sanchez C, Monod J (1960) [Operon: a group of genes with the expression coordinated by an operator]. C R Hebd Seances Acad Sci 250: 1727–1729. Available: http://www.ncbi.nlm.nih.gov/pubmed/14406329. Accessed 3 November 2015.

154. Jacob F, Monod J (1961) Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol 3: 318–356. Available: http://www.ncbi.nlm.nih.gov/pubmed/13718526. Accessed 22 May 2015.

155. Demerec M, Hartman PE (1959) Complex Loci in Microorganisms. Annu Rev Microbiol 13: 377–406. Available: http://www.annualreviews.org/doi/10.1146/annurev.mi.13.100159.002113. Accessed 23 October 2017.

156. Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S (2002) Computational Identification of Operons in Microbial Genomes. Genome Res 12: 1221–1230. Available: http://www.ncbi.nlm.nih.gov/pubmed/12176930. Accessed 23 October 2017.

157. Lawrence JG, Roth JR (1996) Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. Genetics 143: 1843–1860. Available: http://www.genetics.org/content/143/4/1843.abstract?ijkey=0d41823e9bb1780eb0a520a6989cfb93e5064232&keytype2=tf_ipsecsha. Accessed 29 December 2015.

158. Pál C, Hurst LD (2004) Evidence against the selfish operon theory. Trends Genet 20: 232–234. Available:

http://www.sciencedirect.com/science/article/pii/S0168952504000939. Accessed 10 May 2017.

159. Price MN, Huang KH, Arkin AP, Alm EJ (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. Genome Res 15: 809–819. Available: http://www.ncbi.nlm.nih.gov/pubmed/15930492. Accessed 17 July 2017.

160. Rocha EPC (2006) Inference and Analysis of the Relative Stability of Bacterial Chromosomes. Mol Biol Evol 23: 513–522. Available: http://academic.oup.com/mbe/article/23/3/513/1110176/Inference-and-Analysis-of-the-Relative-Stability. Accessed 23 October 2017.

161. Lathe WC, Snel B, Bork P (2000) Gene context conservation of a higher order than operons. Trends Biochem Sci 25: 474–479. Available: http://www.ncbi.nlm.nih.gov/pubmed/11050428. Accessed 23 October 2017.

162. Hershberg R, Yegerlotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulatory network. Trends Genet 21: 138–142. Available: http://www.ncbi.nlm.nih.gov/pubmed/15734572. Accessed 23 October 2017.

163. Warren PB, ten Wolde PR (2004) Statistical Analysis of the Spatial Distribution of Operons in the Transcriptional Regulation Network of Escherichia coli. J Mol Biol 342: 1379–1390. Available: http://www.ncbi.nlm.nih.gov/pubmed/15364567. Accessed 23 October 2017.

164. Korbel JO, Jensen LJ, von Mering C, Bork P (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 22: 911–917. Available: http://www.ncbi.nlm.nih.gov/pubmed/15229555. Accessed 23 October 2017.

165. Cooper S, Helmstetter CE (1968) Chromosome replication and the division cycle of Escherichia coli B/r. J Mol Biol 31: 519–540. Available: http://www.ncbi.nlm.nih.gov/pubmed/4866337. Accessed 23 October 2017.

166. Couturier E, Rocha EPC (2006) Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Mol Microbiol 59: 1506–1518. Available: http://www.ncbi.nlm.nih.gov/pubmed/16468991. Accessed 23 October 2017.

167. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660–665. Available: http://www.ncbi.nlm.nih.gov/pubmed/8676740. Accessed 23 October 2017.

168. Rocha EPC, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34: 377–378. Available: http://www.ncbi.nlm.nih.gov/pubmed/12847524. Accessed 23 October 2017.

169. Rocha EPC, Danchin A (2003) Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res 31: 6570–6577. Available: http://www.ncbi.nlm.nih.gov/pubmed/14602916. Accessed 17 July 2017.

170. Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? Mol Biol Evol 22: 2147–2156. Available: http://www.ncbi.nlm.nih.gov/pubmed/16014871. Accessed 25 June 2016.

171. Fang G, Rocha EPC, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. BMC Genomics 9: 4. Available: http://www.ncbi.nlm.nih.gov/pubmed/18179692. Accessed 25 June 2016.

172. Maynard-Smith, J., R. Burian, S. A. Kauffman, P. Alberch, J. Campbell, B. Goodwin, R. Lande DR and LW (1985) Developmental Constraints and Evolution. Q Rev Biol 60: 265–287. Available: https://www.jstor.org/stable/2828504?seq=1#page_scan_tab_contents.

173. Wagner A (2011) Genotype networks shed light on evolutionary constraints. Trends Ecol Evol 26: 577–584. Available: http://linkinghub.elsevier.com/retrieve/pii/S0169534711001996. Accessed 6 November 2016.

174. Nüsslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in Drosophila. Nature 287: 795–801. Available: http://www.ncbi.nlm.nih.gov/pubmed/6776413. Accessed 6 November 2016.

175. Levitt M (2009) Nature of the protein universe. Proc Natl Acad Sci U S A 106: 11079–11084. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2698892&tool=pmcentrez&rendertype=abstract. Accessed 17 February 2014.

176. Oster GF, Shubin N, Murray JD, Alberch P (1988) Evolution and morphogenetic rules: the shape of the vertebrate limb in ontology and phylogeny. Evolution (N Y) 42: 862–884. Available: http://www.ncbi.nlm.nih.gov/pubmed/28581162. Accessed 13 November 2017.

177. Alberch P, Gale EA (1985) A developmental analysis of an evolutionary trend: digital reduction in amphibians. Evolution (N Y) 39: 8–23. Available: http://www.ncbi.nlm.nih.gov/pubmed/28563634. Accessed 13 November 2017.

178. Stephen Jay Gould (1990) Wonderful Life: The Burgess Shale and the Nature of History. New York: W. W. Norton & Company . Available: https://www.amazon.com/Wonderful-Life-Burgess-Nature-History/dp/039330700X.

179. Lobkovsky AE, Koonin E V (2012) Replaying the tape of life: quantification of the predictability of evolution. Front Genet 3: 246. Available: http://www.ncbi.nlm.nih.gov/pubmed/23226153. Accessed 6 November 2016.

180. Plucain J, Suau A, Cruveiller S, Médigue C, Schneider D, et al. (2016) Contrasting effects of historical contingency on phenotypic and genomic trajectories during a two-step evolution experiment with bacteria. BMC Evol Biol 16: 86. Available: http://www.ncbi.nlm.nih.gov/pubmed/27108090. Accessed 23 October 2017.

181. Bedhomme S, Lafforgue G, Elena SF (2013) Genotypic but not phenotypic historical contingency revealed by viral experimental evolution. BMC Evol Biol 13: 46. Available: http://www.ncbi.nlm.nih.gov/pubmed/23421472. Accessed 23 October 2017.

182. Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia

coli. Proc Natl Acad Sci U S A 105: 7899–7906. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2430337&tool=pmcentrez&rendertype=abstract. Accessed 30 January 2014.

183. Schaper S, Johnston IG, Louis AA (2012) Epistasis can lead to fragmented neutral spaces and contingency in evolution. Proc Biol Sci 279: 1777–1783. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3297459&tool=pmcentrez&rendertype=abstract.

184. Aguirre J, Buldú JM, Stich M, Manrubia SC (2011) Topological structure of the space of phenotypes: the case of RNA neutral networks. PLoS One 6: e26324. Available:

185. Boldhaus G, Klemm K (2010) Regulatory networks and connected components of the neutral space. Eur Phys J B 77: 233–237. Available: http://www.springerlink.com/index/10.1140/epjb/e2010-00176-4.

186. Gould SJ, Vrba ES (1982) Exaptation; a missing term in the science of form. Paleobiology 8: 4–15. Available:

187. Bock WJ (1959) Preadaptation and Multiple Evolutionary Pathways. Evolution (N Y) 13: 194–211. Available: http://www.jstor.org/stable/2405873?origin=crossref&seq=1#page_scan_tab_contents. Accessed 3 January 2016.

188. True JR, Carroll SB (2002) Gene co-option in physiological and morphological evolution. Annu Rev Cell Dev Biol 18: 53–80. Available: http://www.ncbi.nlm.nih.gov/pubmed/12142278. Accessed 3 December 2015.

189. Zákány J, Duboule D (1999) Hox genes in digit development and evolution. Cell Tissue Res 296: 19–25. Available: http://www.ncbi.nlm.nih.gov/pubmed/10199961. Accessed 3 January 2016.

190. Keys DN, Lewis DL, Selegue JE, Pearson BJ, Goodrich L V, et al. (1999) Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. Science 283: 532–534. Available: http://www.ncbi.nlm.nih.gov/pubmed/9915699. Accessed 3 January 2016.

191. Colleen Farmer (1997) Did Lungs and the Intracardiac Shunt Evolve to Oxygenate the Heart in Vertebrates? on JSTOR. Paleobiology 23: 358–372. Available:

192. Pievani T, Serrelli E (2011) Exaptation in human evolution: how to test adaptive vs exaptive evolutionary hypotheses. J Anthropol Sci 89: 9–23. Available: http://www.ncbi.nlm.nih.gov/pubmed/21757789. Accessed 3 January 2016.

193. Tomarev SI, Piatigorsky J (1996) Lens crystallins of invertebrates--diversity and recruitment from detoxification enzymes and novel proteins. Eur J Biochem 235: 449–465. Available: http://www.ncbi.nlm.nih.gov/pubmed/8654388. Accessed 3 January 2016.

194. Barve A, Wagner A (2013) A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500: 203–206. Available: http://dx.doi.org/10.1038/nature12301. Accessed 11 July 2014.

# Chapter 2:

# Phenotypic innovation through recombination in genome-scale metabolic networks

Sayed-Rzgar Hosseini, Olivier C Martin and Andreas Wagner

The content of this chapter has been published as:

Hosseini, S.-R., O. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc. R. Soc. B.
DOI: 10.1098/rspb.2016.1536

## 2.1. Abstract

Recombination is an important source of metabolic innovation, especially in prokaryotes, which have evolved the ability to survive on many different sources of chemical elements and energy. Metabolic systems have a well-understood genotype-phenotype relationship, which permits a quantitative and biochemically principled understanding of how recombination creates novel phenotypes. Here we investigate the power of recombination to create genome-scale metabolic reaction networks that enable an organism to survive in new chemical environments. To this end, we use flux balance analysis, an experimentally validated computational method that can predict metabolic phenotypes from metabolic genotypes. We show that recombination is much more likely to create novel metabolic abilities than random changes in chemical reactions of a metabolic network. We also find that phenotypic innovation is more likely when recombination occurs between parents that are genetically closely related, phenotypically highly diverse, and viable on few rather than many carbon sources. Survival on a new carbon source preferentially involves reactions that are superessential, that is, essential in many metabolic networks. We validate our observations with data from 61 reconstructed prokaryotic metabolic networks. Our systematic and quantitative analysis of metabolic systems helps understand how recombination creates innovation.

## 2.2. Introduction

Organisms that reproduce sexually and recombine their DNA bear high evolutionary costs, among them the disruption of well-adapted phenotypes caused by recombination. Nonetheless, recombination is common in nature. This paradox has spurred many efforts to resolve it [1–3].

From a genetic perspective, the major benefit and costs of recombination are similar in kind, because recombination can create well-adapted phenotypes just as it can destroy them. Recombination has the ability to join beneficial mutations from two organisms or molecules [4,5], and thus speed up adaptive evolution [6]. In recent years, experimental evidence obtained through DNA shuffling experiments has made clear how great this benefit of recombination can be in creating proteins with novel phenotypes [7]. Many experimental studies have confirmed the power of recombination in generating genes, pathways, and genomes with novel features [8–11].

Because recombination can involve large-scale genotypic change, its power to disrupt existing well adapted phenotypes may also be large. However, recent directed evolution experiments and computational studies based on model transcriptional gene regulatory circuits and lattice proteins, suggest that random mutations can be more likely to disrupt existing phenotypes than recombination [12–14].

To understand the relative costs and benefits of recombination, one needs to understand how genotypic change causes phenotypic change, but this is difficult even for well-studied systems like proteins. For example, measuring the phenotypic effects of genetic changes engineered into proteins is laborious, and our ability to predict altered protein functions from altered amino acid sequences computationally is very limited. However, in one class of biological systems, the complex and genetically encoded chemical reaction network of metabolism, we understand genotype-phenotype relationships better [15–18]. The reason is that computational tools such as flux balance analysis (FBA) provide a means to predict metabolic phenotypes – the ability of an organism to survive on specific nutrients – from information about metabolic genotypes, i.e., the collection of chemical reactions that a metabolic reaction network is able to catalyze [19]. What is more, FBA-based qualitative

predictions of metabolic phenotypes are in good agreement with experimental data [20]. We here use FBA to quantify and understand the disruptive and creative effects of recombination on the biochemistry of metabolic systems. We will focus especially on metabolic innovation, the ability of recombination to create metabolic networks that are able to survive on new sources of carbon and energy. In doing so, we represent metabolic genotypes not as DNA sequences but as sets of metabolic reactions that can be altered through recombination, a common approach in metabolic systems biology [15–18,21–25].

Metabolism is an ideal system to study innovation, especially for microorganisms, because the prokaryotic world is rife with examples of metabolic innovations. For instance, microorganisms have acquired the capability to extract energy from a bewildering variety of non-natural and even toxic substances [26–29]. Microbial isolates from pristine soils not only have acquired resistance to a wide range of antibiotics, but they can even use some of these molecules as food [30]. Halophilic bacteria can tolerate high salt concentration by synthesizing novel molecules such as ectoine or glycine betaine [31].

In eukaryotes, recombination occurs during meiosis and is thus linked to reproduction. It involves parents that are usually genetically similar and belong to the same population and species. In contrast, prokaryotic recombination is not usually linked to reproduction. It occurs via horizontal gene transfer [32], whose incidence is large and greater than that of point mutations [33–35]. It changes the organization and gene content of genomes on short evolutionary time scales [32,36,37], and can involve very distantly related organisms [38,39]. Although horizontal gene transfer adds genes from a donor to a recipient, incorporating such genes into the recipient genome relies on DNA rearrangements that can also delete resident genes [40]. More generally, the majority of newly acquired genes obtained via horizontal gene transfer reside in the genome only for short amounts of time [41], and prokaryotic genomes show a bias towards DNA deletions [42]. Motivated by these observations, we here model prokaryotic recombination as a process where the transfer of biochemical reactions from a donor to a recipient is accompanied by concurrent deletion of reactions from the recipient.

Our work builds on an approach that we developed previously to study *typical* properties of a metabolic network with a given phenotype – the ability to survive on a

given set of carbon and energy sources. These are properties that are independent of any one organism such as *E. coli* [18,23–25,43,44]. The method explores a vast space of possible metabolic genotypes to create a random sample of metabolic networks that are viable on specific carbon sources such as glucose. We here use this approach to create pairs of ”parental” metabolic networks with well-defined genotypes and phenotypes. We ask how likely it is that recombination between these parents (i) disrupts their metabolic phenotypes, and (ii) creates novel, innovative metabolic networks that can survive on at least one novel sources of carbon and energy, among 50 different such sources we consider. We validate our observations with data from 61 prokaryotic genome-scale metabolic networks.

Our observations show that recombination creates more metabolic innovations than an equivalent amount of random change in a metabolic network's reaction complement – our model's representation of random mutation. At the same time recombination is no more disruptive than random change. Importantly, the innovative power of recombination increases with the phenotypic diversity of the parents. In contrast, it decreases with their genotypic diversity and with their phenotypic complexity (the number of carbon sources on which they are viable). Moreover, we find that a class of metabolic reactions that we refer to as superessential plays an important role in metabolic innovation [45].

## 2.3. Results

### *(a) Recombination causes more metabolic innovations than random change*

To quantify the power of recombination in creating novel phenotypes in metabolic systems, we created 1000 donor-recipient pairs of random viable metabolic networks with a fixed metabolic genotypic distance of $D=100$ reactions. (Genomic data shows that bacteria at this or greater metabolic divergence often recombine successfully: see electronic supplementary material text S4, and figure S4). For each of these pairs, we generated 1000 recombinant offspring by recombining a given number $n$ of reactions (See Methods). We quantified the incidence of metabolic innovation as the fraction ($f_{innov}$) of offspring retaining viability on glucose that also gain viability on at least one additional carbon source. Moreover, we compared recombination's effects on innovation with those of an equivalent amount of random change (“mutation”, Methods). That is, we computed the fraction of innovative offspring ($f_{innov}$) for

metabolic networks with a number of random reaction deletions or additions equivalent to that caused by recombination.

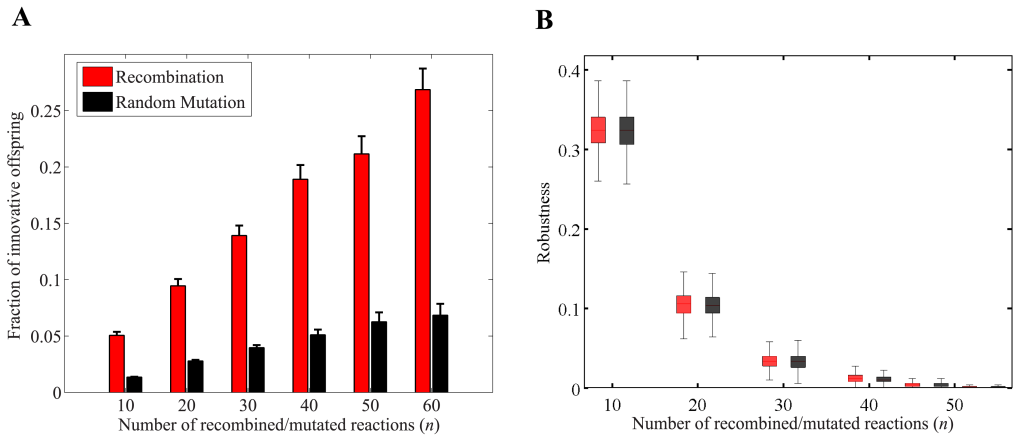



**Figure 1: Cost-benefit relationship of recombination as compared with random reaction change ("mutation"). (a)** Mean (bars) and standard deviation (vertical lines) of the fraction of innovative offspring ($f_{innov}$) among all offspring retaining viability on glucose that are generated by recombination (red) and random reaction change (black) , as a function of the number of reaction changes ($n$, $x$-axis). **(b)** Recombinational robustness (red), that is, the fraction of recombinant offspring retaining viablity on glucose, and mutational robustness, that is, the fraction of mutant offspring that can retain viability on glucose, as a function of the number of reaction changes ($n$, $x$-axis). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

Figure 1a shows the mean and standard deviation of $f_{innov}$ as a function of $n$. Two patterns are germane. First, the fraction of innovative offspring ($f_{innov}$) is consistently greater for recombination (figure 1a, red) than for random change (black). For example, for recombination events involving $n=10$ reactions, 5 percent of the viable recombinant offspring could gain viability on at least one novel carbon source on average. In contrast, for random change involving $n=10$ reactions, this fraction is more than 4 times smaller, consistently below 1.3 percent. Second, $f_{innov}$ increases with the number of reactions exchanged through recombination. Our observations are robust to an alternative approach to generating metabolic network pairs with a fixed $D=100$, and to the choice of an alternative primary carbon source (See methods and electronic supplementary material, figures S3a, S3b, and S3c). In sum, recombination produces a higher incidence of innovation. The diversity of metabolic innovations it produces is similar to those produced by random change (electronic supplementary figure S5).

Next we quantified the cost of recombination, i.e., its power in disrupting existing phenotypes. To this end, we measured recombinational robustness, the fraction of a parental pair's recombinant offspring that retains viability on glucose. Figure 1b (red) shows the recombinational robustness of the 1000 donor-recipient pairs as a function of the number *n* of exchanged reactions. We also wanted to compare this recombinational robustness with our model's equivalent of mutational robustness, i.e., robustness to a comparable amount of random additions and deletions of reactions from a metabolic network. Figure 1b shows that recombinational robustness is not lower than robustness to random reaction change, regardless of the number (*n*) of altered reactions. Again, an alternative approach to generating metabolic network pairs with a fixed *D=100,* or an alternative primary carbon source yields similar observations (See methods and electronic supplementary material, figures S3d, S3e, and S3f). We also note that innovative offspring grow even faster than parental metabolic networks on the carbon source where their parents were viable (electronic supplementary material, figure S6). Therefore, recombination can cause more innovation than random change, but it does not incur higher costs.

***(b) Superessential reactions play an important role in metabolic innovation***

To understand the specific reaction changes associated with metabolic innovation, we next analyzed all 8171 recombinant offspring with metabolic innovations that our analysis had identified. In most of them, only one of the multiple reactions added in a recombination event was responsible for gaining viability on a novel carbon source. For example, among three recombinant offspring that had independently gained viability on acetate, all three had gained the phosphoglycerate kinase reaction. Likewise, the five recombinant offspring that had independently gained viability on pyruvate achieved this gain through addition of the ribulose 5-phosphate 3-epimerase reaction in the pentose phosphate pathway. More generally in 98.91 percent of innovative offspring, a single reaction accounted for the innovation. An example of the 89 (1.09 percent) instances where multiple reaction additions are responsible for an innovation is the newly acquired viability on the carbon source trehalose. It was caused by simultaneous addition of reactions catalyzed by trehalose 6-phosphate phosphorylase and 2-dehydro-3-deoxy-phosphogluconate aldolase.

The reactions that cause viability on new carbon sources come from a relatively small subset of the "universe" of 5906 reactions (see Methods). Specifically, only 19 reactions among the 5906 reactions are responsible for gaining viability on new carbon sources in the majority (53%, 4430) of the 8171 innovative offspring. The remaining 47% of innovations are caused by only 147 other reactions. What is more, these reactions tend to share a property that we refer to as their superessentiality [45]. The superessentiality index ($I_{SE}$) of a metabolic reaction denotes the fraction of metabolic networks in which this reaction is essential for viability on carbon source C [45]. It can be computed from randomly sampled metabolic networks viable on that carbon source. The greater a reaction's superessentiality index ($I_{SE}$) the larger is also the number of bacterial genomes encoding this reaction [45]. Reactions with $I_{SE} > 0.5$ are essential for viability in the majority of metabolic networks where they occur.

Reactions that cause viability on a new carbon source tend to have a higher superessentiality index than those, which rarely or never cause viability on new carbon sources (Figure S7a). Examples include the ribulose-5-phosphate 3-epimerase reaction, which causes viability on new carbon sources in 731 innovative offspring, and has a superessentiality index of $I_{SE}=0.9714$. They also include ribose-5-phosphate isomerase ($I_{SE}=0.9530$), which causes metabolic innovation in 677 offspring. More generally, we observed (i) a significant correlation between the superessentiality index and the number of innovations a reaction causes (figure S7b), and (ii) that the fraction of innovative offspring ($f_{innov}$) of metabolic network pairs increases with $f_{super}$, i.e., with the fraction of reactions with $I_{SE} > 0.5$ (electronic supplementary material, figure S7c; Pearson's $r=0.18$, $P<10^{-8}$).

### *(c) Genotypically more similar parental metabolic networks are more likely to generate metabolically innovative offspring*

Thus far we examined the effects of recombination on metabolic innovation and robustness among parental metabolic networks with a fixed genotypic distance $D$. This distance, however, could have an influence on the effect of recombination. For example, recombination among more distant parents could lead to a smaller incidence of metabolic innovation, because fewer reactions "imported" from a distant metabolic network may integrate productively into the resident metabolic network. To find out whether this is the case, we varied the distance among recombining parents between

*D=100* and *D=1500*. We did not examine greater distances, because according to available data on the rate of successful recombination among prokaryotic species, this rate becomes negligible at such large distances (electronic supplementary material, text S4). However, as a reference point for including parents with the maximal genotype distance ($D_{max}$), we analyzed random metabolic changes, where new reactions are added not from another parental metabolic network but from the (maximally diverse) reaction universe. Specifically, for each value of $D$, we created 1000 random pairs of parental metabolic networks, and from each pair we formed 1000 recombinant offspring by recombining a fixed number $n$ of randomly chosen reactions (see Methods).

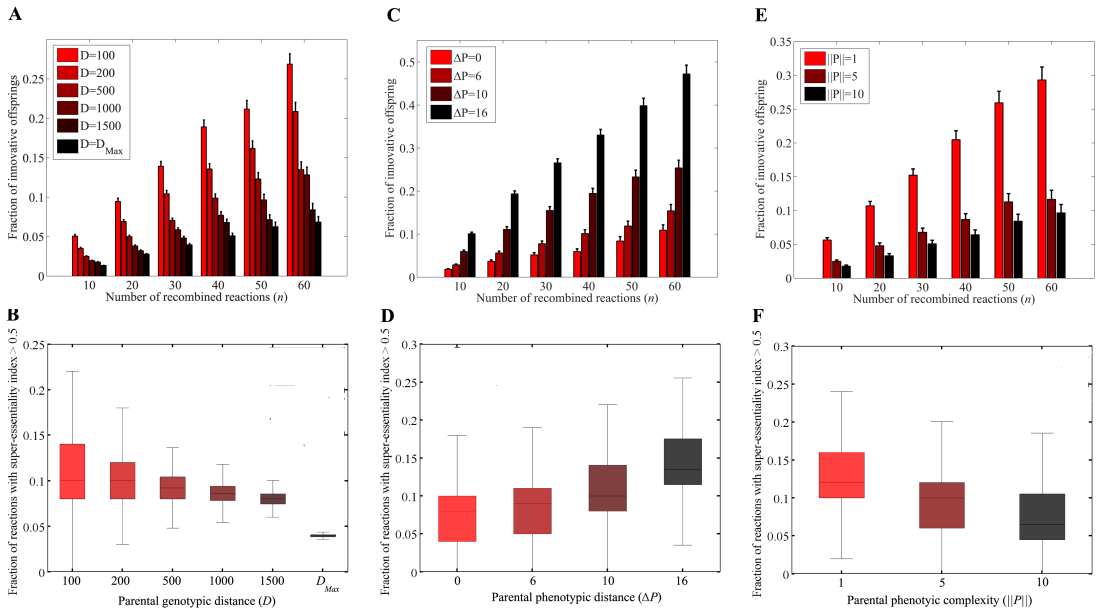



**Figure 2: Effect of parental genotypic and phenotypic features on the incidence of innovative offspring.** The vertical axes show mean (bar) and standard deviation (vertical line) of the fraction of innovative offspring ($f_{innov}$), generated by recombination between parental metabolic networks with **(a)** genotypic distance ($D$), **(c)** phenotypic distance ($\Delta P$), and **(e)** phenotypic complexity (||P||), where $D$, $\Delta P$, and $||p||$ are color-coded according to their corresponding legend. The horizontal axis shows the number of recombined reactions ($n$). In panels (b), (d), and (f) , the vertical axes show the fraction of reactions with superessentiality index exceeding 0.5 ($f_{super}$, $x$-axis) among reactions that can potentially be transferred from the parental donor to the recipient metabolic network. The horizontal axes show in **(b)** genotypic distance ($D$), **(d)** phenotypic distance ($\Delta P$), and **(f)** phenotypic complexity ($||P||$). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

Figure 2a shows that for any given number $n$ of recombining reactions, the fraction of innovative offspring ($f_{innov}$) decreases with increasing parental genotype distance. In other words, the more distant two recombining parents are, the smaller the likelihood that their offspring can survive on novel carbon sources. Although this relationship could be the result of an increasing fraction of inviable offspring when distantly related parents recombine, electronic supplementary material, figure S8 shows that this is not the case. These observations still hold if we require parental metabolic networks to be viable on acetate instead of glucose (electronic supplementary material, figures S9a and S9b).

We also examined how the fraction of reactions with $I_{SE} > 0.5$ ($f_{super}$) that can potentially be transferred from donor to the recipient metabolic network changes with genotypic distance. Both the median and the variance of $f_{super}$ decreases with increasing $D$ (Figures 2b, and electronic supplementary material, figure S9c). This observation further supports the importance of highly superessential reactions for metabolic innovation.

Moreover, we showed that parental metabolic networks with higher phenotypic diversity have greater potential to create innovative offspring (electronic supplementary material, texts S5 and S6, and figures 2c, 2d, S10, and S11). In addition, we also found that parental metabolic networks with more reactions (higher genotypic complexity), and those viable on fewer carbon sources (lower phenotypic complexity) are more likely to generate innovative offspring (electronic supplementary material, texts S7 and S8, and figures 2e, 2f, S12, S13, and S14).

### *(d) Recombination in prokaryotic metabolic networks has similar innovation potential as in randomly sampled metabolic networks*

While sampling viable metabolic networks from a metabolic genotype space permitted us to control parameters such as phenotypic and genotypic diversity, this analysis also has limitations. For example, it neglects the potential influence of gene linkage on metabolic innovation by recombination, because it is based on the exchange of biochemical reactions rather than genes. We thus wanted to validate our observations with genome-scale metabolic networks of prokaryotic organisms. To this end, we used genome-scale metabolic networks from 61 bacterial species from the BiGG database [50], which differ in both their genotypes and phenotypes, i.e.,

their viability on 137 different carbon sources (See Methods). We first determined the innovation potential of each of the 3660 possible pairs that can be formed from these metabolic networks, by asking whether the union of a pair's reaction sets confers viability on a carbon source on which neither member of the pair is viable. This was the case for 1126 pairs (30.77%).

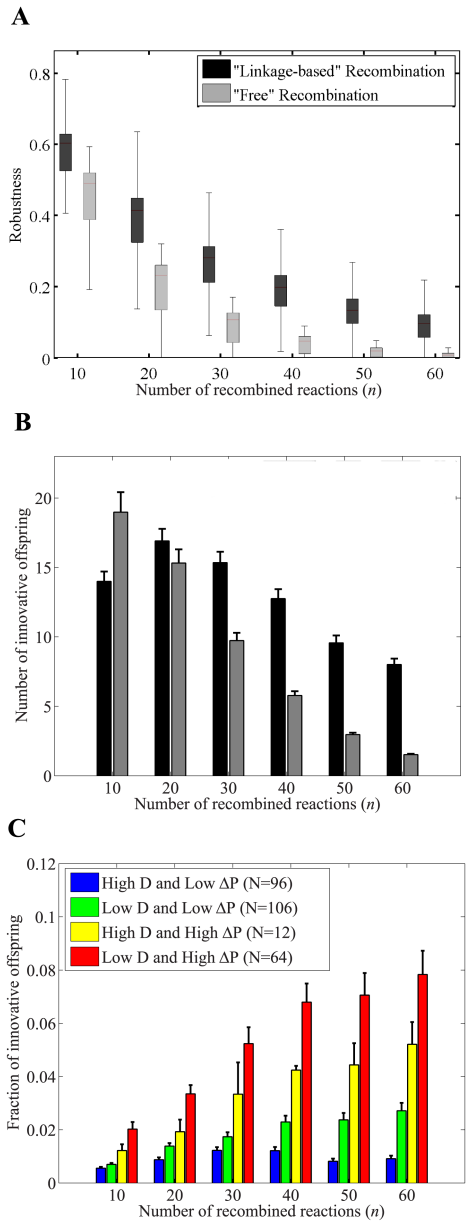



**Figure 3: Recombination between prokaryotic metabolic networks shows the same innovation potential as in randomly sampled metabolic networks. (a)** Robustness to recombination, i.e., the fraction of recombinant offspring retaining viability on at least one parental carbon source, and **(b)** Mean (bar), and standard error (vertical line) of the number of innovative offspring among 1000 recombinant offspring (y-axis), as a function of the number of recombined reactions (x-axis). Offspring were generated by i) linkage-based recombination between prokaryotic metabolic networks (black), and ii) free recombination between prokaryotic metabolic networks (gray). **(c)** The vertical axis shows the mean (bar) and standard deviation (vertical line) of the fraction of innovative offspring ($f_{innov}$) generated by linkage-based recombination between prokaryotic parental metabolic networks with i) high genotypic distance ($D>40$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (blue, $N=96$ parental pairs), ii) low genotypic distance ($D<30$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (green, $N=106$ parental pairs), iii) high genotypic distance ($D>40$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (yellow, $N=12$ parental pairs), and iv) low genotypic distance ($D<30$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (red, $N=64$ parental pairs).

For each of these 1126 metabolic network pairs, we generated 1000 recombinant offspring through a procedure we refer to as "linkage-based" recombination, which maps reactions onto genes, and recombines random stretches of DNA whose length is chosen such that a given number of *n* reactions is altered in the offspring's metabolic network (see Methods). For the purpose of comparison, we also used a complementary "free recombination" procedure, which disregards linkage and creates recombinant offspring by altering a given number *n* of reactions in the recipient, just as we had done for randomly sampled viable networks. When analyzing robustness to recombination, we found that linkage-based recombination is much more likely to preserve viability than free recombination (Figures 3a, and electronic supplementary material, figure S15a). In other words, the linear organization of metabolic genes in the genome facilitates robustness to recombination.

The fraction of innovative offspring is somewhat lower under linkage-based recombination (electronic supplementary material, figure S15b). However, a higher overall fraction of viable offspring (figure 3a) results in substantially higher total number of innovative offspring under linkage-based recombination than free recombination, (especially for higher numbers of recombined reactions (*n*), (figure 3b)). Therefore, higher robustness to recombination in prokaryotic metabolic networks results in a higher potential for metabolic innovation.

In a majority of the 1126 prokaryotic metabolic network pairs where metabolic innovation is possible (854 of 1126, or 75.84%), the addition of a single reaction from donor to the recipient was sufficient to gain viability on a new carbon source, just as we had observed for randomly sampled metabolic networks. Also, only a small fraction (106) of the 3404 reactions that occurred in these 61 metabolic networks caused metabolic innovation. And just as in randomly sampled metabolic networks, reactions that cause innovation are more often essential (higher $I_{SE}$ [45]) than those that do not cause innovation (electronic supplementary material, figure S15c). In addition, the fraction of innovative offspring ($f_{innov}$) increases with the fraction of reactions with superessentiality index $I_{SE}$ larger than 0.5 (electronic supplementary material, figure S15d; Pearson's $r=0.13$, $P<10^{-5}$).

Finally, we showed that parental prokaryotic metabolic networks with low genotypic distance, high phenotypic distance and low phenotypic complexity are more likely to generate innovative offspring (figure 3c, and electronic supplementary material, text S9, and figures S16 and S17a), just as they did in randomly sampled viable metabolic networks. Moreover, genotypic distance, phenotypic complexity, and phenotypic diversity do not strongly influence recombinational robustness, just as they did not for randomly sampled metabolic networks (electronic supplementary material, text S9, and figures S16, S17b, and S17c).

In sum, our observations show that recombination in prokaryotic metabolic networks resembles those in randomly sampled metabolic networks in its innovation potential, and in the mechanisms by which it causes metabolic innovation.

## 2.4. Discussion

Recombination is a major force behind many innovations in biological systems [12–14,53,54]. Here we studied its innovation potential in metabolic systems, where innovations enable organisms to survive on novel sources of energy and chemical elements. To do so, we computationally recombined biochemical reactions among thousands of metabolic network pairs that are viable on specific carbon sources. We sampled most of these from a vast space of such networks with a Markov Chain Monte Carlo technique, but also validated our observations by analyzing 61 prokaryotic metabolic networks.

We found that recombination provides greater benefit – a greater number of new metabolic abilities – at no greater cost in terms of viability loss than an equivalent amount of random mutation, modeled as random alterations in a metabolic network's complement of chemical reactions. Metabolic innovation is more likely when recombination occurs between parents that are genetically closely related, phenotypically highly diverse, and viable on few rather than many carbon sources. Survival on a new carbon source preferentially involves reactions that are highly super-essential, that is, they are required in most metabolic networks viable on this carbon source.

One well-studied facilitator of evolutionary adaptation and innovation is the robustness of a biological system to genetic change [55,56]. Robustness can facilitate

a population's exploration of a genotype space, and thus accelerate the origin of novel phenotypes [57]. Our observations support this positive role of robustness. Specifically, the frequently higher robustness of prokaryotic metabolic networks under linkage-based recombination results in a higher total number of innovative offspring (Figures 3a and 3b). Moreover, larger metabolic networks are more robust to recombination, and they are more likely to create metabolic innovations (electronic supplementary material, figure S14). We also showed, however, that robustness is not the only factor affecting innovation by recombination. For example, parental genotypic distance, phenotypic diversity and complexity impact innovation without influencing robustness, so they modulate the incidence of innovation independent of robustness. The ability to study these factors in isolation is a key advantage of our computational approach.

Systemic properties like robustness and parental diversity are not the only factors influencing innovation by recombination. One property of individual reactions – superessentiality – is at least as important. Multiple reactions may be transferred in a recombination event, but in the vast majority of such events, only the addition of a single reaction causes innovation, and this reaction is often highly superessential. What is more, reaction superessentiality can help explain multiple systemic patterns in our data, e.g., that phenotypically diverse parents have greater innovation potential (see electronic supplementary material, text S10, and figure S18). That being said, exceptions to the importance of superessentiality exist, where innovations are caused by reactions that are rarely essential. An extreme example is adenyl cyclase, which catalyses the conversion of ATP to 3',5'-cyclic AMP and diphosphate. It is not essential for viability on any of the 10000 randomly sampled metabolic networks ($I\text{-}_{SE}=0$), yet it is responsible for metabolic innovation in 10 out of 8171 innovative offspring. Relatedly, we found 3 examples of reactions that were blocked (i.e. inactive, with zero flux) in the donor metabolic network that caused innovation after being added to a recipient metabolic network. However, the innovation potential of such inactive reactions is small compared to active or highly superessential reactions (electronic supplementary material, figure S2).

Our approach of using randomly sampled viable metabolic networks has several advantages, most notably that we can arrive at general conclusions that go beyond any one organism, and that we can control important quantities such as parental

genotypic diversity. However, it also has several limitations. First, our computational analysis is based on FBA, which neglects regulatory constraints that can arise through suboptimal enzyme expression. However, as we discussed in more detail in online supplementary material, text S1d, this limitation is not likely to affect our main observations.

Second, our approach ignores the linkage of related metabolic genes on chromosomes, for example in operons [58]. Although on long evolutionary time scales operons often break up and reform [59,60], functionally related genes tend to be linked. Randomly sampled metabolic networks may contain combinations of reactions that are not found in any known organism, so that we cannot meaningfully assign linkage patterns to them. Third, our specification of a metabolic genotype represents this genotype on the level of the reaction rather than that of a gene and considers only the presence or absence of metabolic reactions. Although widely used [15–18,21–25], this representation neglects potentially important information, among them myriad mechanistic details of DNA recombination. Perhaps even more importantly, it also neglects that some reactions are catalyzed by multiple enzymes [61], and that some enzymes catalyze multiple reactions [62–64].

We were able to mitigate the last two limitations by comparing our observations with those obtained from 61 curated prokaryotic metabolic networks, where gene-reaction maps and linkage information is available for metabolic genes. The incidence of metabolic innovation among hundreds of pairs of these prokaryotic metabolic networks shows the same patterns as our randomly sampled viable metabolic networks. In addition, these metabolic networks revealed an additional intriguing pattern, their increased metabolic robustness to recombination. Essential genes in general are known to be clustered in prokaryotic genomes [65,66], which may increase a genome's robustness to large-scale gene deletions. However, the evolutionary forces shaping the organization of metabolic genes are not well known, and call for further research.

Recombination and DNA mutations, such as point mutations and gene duplications, play complementary roles in creating metabolic innovation. In an evolving population, recombination cannot be effective unless mutation has created diversity beforehand. Mutations introduce novel parts (enzymes, reactions), and recombination

creates novel combinations of these parts (metabolic pathways). Our results demonstrate the power of this combinatorial principle for metabolic innovation. One source of this power is that recombination shuffles system parts that have been "pre-tested" in evolution, because they form part of a viable metabolic network. This is also why superessential reactions are important for innovation, and why robustness facilitates innovation.

## 2.5. Methods

### *(a) Genome-scale metabolic networks and their phenotypic representations*

A metabolism is a network of enzyme-catalyzed biochemical reactions. Each such metabolic network contains a subset of the "reaction universe" of all biochemical reactions that take place in the biosphere. We have curated a representation of this universe, which comprises 5906 reactions and is based on current metabolic knowledge (for more details, see electronic supplementary material text S1a and S1b) [18,46–49]. We represent an organism's metabolic genotype as a binary vector of length 5906. Each entry of this vector corresponds to a given reaction in the universe, and is equal to one if the corresponding reaction is present in the network, and zero otherwise. Thus, each genotype can be thought of as a single member of a vast space of all possible metabolic networks, which contains $2^{5906}$ distinct genotypes. We define the phenotype of a given metabolic genotype based on its viability on 50 distinct minimal environments that differ only in the carbon source (electronic supplementary material text S1c). We consider a genotype *viable* on a given carbon source, if it can produce all the essential biomass precursors from the given carbon source, and we use Flux Balance Analysis (FBA, See electronic supplementary material text S1d) to determine viability [19].

### *(b) Generation of random metabolic networks*

We here employ a previously described *in silico* process which relies on Markov Chain Monte Carlo (MCMC) random walks to generate metabolic networks comprising random sets of reactions that are viable on a given carbon source (electronic supplementary material text S1e) [18,23]. This procedure can produce metabolic networks that are sampled uniformly from the set of all metabolic networks viable on a given carbon source.

### *(c) Generation of parental metabolic network pairs*

Our analyses required us to recombine pairs of "parental" metabolic networks with particular features, such as (i) their genotypic distance ($D$), defined as the number of reactions differing between the parents, (ii) their phenotypic complexity ($||P||$), that is, the number of carbon sources on which they are viable, (iii) their phenotypic distance ($\Delta P$), that is, the number of carbon sources on which only one but not the other member of a parental pair is viable, and (iv) their genotypic complexity ($||G||$, or metabolic network size), defined as the number of reactions in each metabolic network. We used simultaneous genotype-converging MCMC random walks to generate pairs of metabolic networks with the features described above (See electronic supplementary material texts S1f, and S2).

### *(d) Modeling recombination and mutation in metabolic networks*

To implement recombination for each parental metabolic network pair, we generated 1000 recombinant offspring by (i) adding to the recipient metabolic network a given number $n/2$ of randomly chosen reactions that were present in the donor and absent in the recipient, followed by (ii) deleting $n/2$ reactions randomly chosen from the recipient. Thus, the total number of reactions changed by a recombination event in the recipient is equal to $n$ (for more details, see electronic supplementary material texts S1g, S3, and figures S1, S2 and S3). We then quantify the incidence of metabolic innovation as the fraction ($f_{innov}$) of offspring retaining viability on parental carbon sources that also gain viability on at least one additional carbon source.

To implement an amount of random change – our model's equivalent of "mutation"– in a metabolic network comparable to the same amount of recombinational change for a given $n$, we created a network's "mutational" offspring by adding $n/2$ randomly chosen reactions from the reaction universe, and deleting $n/2$ randomly chosen reactions from the network. Note that the $n/2$ reactions added to recombinant offspring are chosen randomly from another viable network (the donor), whereas in mutation they are taken from the whole reaction universe.

To validate our model's results, we also analyzed the metabolic networks of 61 prokaryotes obtained from the BIGG database [50], using the R package Sybil [51]. In these networks, we incorporated information about the linkage of the genes encoding metabolic reactions. To this end, we used the gene-reaction association

rules defined in the BiGG database for each organism (in MAT files, grRules) [50], and ordered the genes in each organism based on their genomic position, as obtained from the RefSeq microbial genome database [52] (for more details see electronic supplementary material text S1h).

## 2.6. References


1. Smith, J. M. 1978 *The Evolution of Sex*. Cambridge: Cambridge University Press.
2. Otto, S. P. & Lenormand, T. 2002 Resolving the paradox of sex and recombination. *Nat. Rev. Genet.* 3, 252–61.
3. Hartfield, M. & Keightley, P. D. 2012 Current hypotheses for the evolution of sex and recombination. *Integr. Zool.* 7, 192–209.
4. Fisher, R. A. 1930 *The Genetical Theory of Natural Selection: A Complete Variorum Edition*. OUP Oxford.
5. Muller, H. J. 1932 Some Genetic Aspects of Sex. *Am. Nat.* , 118–138.
6. Keightley, P. D. & Otto, S. P. 2006 Interference among deleterious mutations favours sex and recombination in finite populations. *Nature* 443, 89–92.
7. Stemmer, W. P. 1994 DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci.* 91, 10747–10751.
8. Zhang, Y.-X., Perry, K., Vinci, V. A., Powell, K., Stemmer, W. P. C. & del Cardayré, S. B. 2002 Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature* 415, 644–6.
9. Crameri, A., Dawes, G., Rodriguez, E., Silver, S. & Stemmer, W. P. 1997 Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nat. Biotechnol.* 15, 436–8.
10. Chang, C. C., Chen, T. T., Cox, B. W., Dawes, G. N., Stemmer, W. P., Punnonen, J. & Patten, P. A. 1999 Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* 17, 793–7.
11. Ness, J. E., Welch, M., Giver, L., Bueno, M., Cherry, J. R., Borchert, T. V, Stemmer, W. P. & Minshull, J. 1999 DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* 17, 893–6.
12. Martin, O. C. & Wagner, A. 2009 Effects of recombination on complex regulatory circuits. *Genetics* 183, 673–84, 1SI–8SI.
13. Cui, Y., Wong, W. H., Bornberg-Bauer, E. & Chan, H. S. 2002 Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 809–14.
14. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. 2005 On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5380–5.

15. Feist, A. M. & Palsson, B. Ø. 2008 The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* 26, 659–67.

16. McCloskey, D., Palsson, B. Ø. & Feist, A. M. 2013 Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* 9, 661.

17. Lewis, N. E., Nagarajan, H. & Palsson, B. O. 2012 Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305.

18. Matias Rodrigues, J. F. & Wagner, A. 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* 5, e1000613.

19. Orth, J. D., Thiele, I. & Palsson, B. Ø. 2010 What is flux balance analysis? *Nat. Biotechnol.* 28, 245–8.

20. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. 2001 In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* 19, 125–30.

21. Edwards, J. S. & Palsson, B. O. 2000 The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5528–33.

22. Edwards, J. S. & Palsson, B. O. 1999 Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* 274, 17410–6.

23. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. 2010 Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* 4, 30.

24. Barve, A., Hosseini, S.-R., Martin, O. C. & Wagner, A. 2014 Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. *BMC Syst. Biol.* 8, 48.

25. Hosseini, S.-R., Barve, A. & Wagner, A. 2015 Exhaustive Analysis of a Genotype Space Comprising 1015 Central Carbon Metabolisms Reveals an Organization Conducive to Metabolic Innovation. *PLoS Comput. Biol.* 11, e1004329.

26. Copley, S. D. 2000 Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem. Sci.* 25, 261–5.

27. Rehmann, L. & Daugulis, A. J. 2008 Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. *Biotechnol. Bioeng.* 99, 521–8.

28. Van der Meer JR, Werlen, C., Nishino, S. & Spain, J. 1998 Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. *Appl. Environ. Microbiol.* 64, 4185–93.

29. Cline, R. E., Hill, R. H., Phillips, D. L. & Needham, L. L. In press. Pentachlorophenol measurements in body fluids of people in log homes and workplaces. *Arch. Environ. Contam. Toxicol.* 18, 475–81.

30. Dantas, G., Sommer, M. O. A., Oluwasegun, R. D. & Church, G. M. 2008 Bacteria subsisting on antibiotics. *Science* 320, 100–3.

31. Detkova, E. N. & Boltyanskaya, Y. V. 2007 Osmoadaptation of haloalkaliphilic bacteria: Role of osmoregulators and their possible practical application. *Microbiology* 76, 511–522.

32. Thomas, C. M. & Nielsen, K. M. 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–21.

33. Guttman, D. S. & Dykhuizen, D. E. 1994 Clonal divergence in Escherichia coli as a result of recombination, not mutation. *Science* 266, 1380–3.

34. Feil, E. J. et al. 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U. S. A.* 98, 182–7.

35. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. 2005 Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol. Evol.* 22, 2354–61.

36. Ochman, H., Lawrence, J. G. & Groisman, E. A. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304.

37. Pál, C., Papp, B. & Lercher, M. J. 2005 Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37, 1372–5.

38. Fraser, C., Hanage, W. P. & Spratt, B. G. 2007 Recombination and the nature of bacterial speciation. *Science* 315, 476–80.

39. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. 2000 Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J. Bacteriol.* 182, 1016–23.

40. Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. & Rehrauer, W. M. 1994 Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.* 58, 401–65.

41. Kuo, C.-H. & Ochman, H. 2009 The fate of new bacterial genes. *FEMS Microbiol. Rev.* 33, 38–43.

42. Mira, A., Ochman, H. & Moran, N. A. 2001 Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–96.

43. Matias Rodrigues, J. F. & Wagner, A. 2011 Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst. Biol.* 5, 39.

44. Wagner, A., Andriasyan, V. & Barve, A. 2014 The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. *J. Mol. Biochem.* 3.

45. Barve, A., Rodrigues, J. F. M. & Wagner, A. 2012 Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1121–30.

46. Goto, S., Nishioka, T. & Kanehisa, M. 2000 LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* 28, 380–2.

47. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. 2010 KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–60.

48. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. 2006 From

genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–7.

49. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. & Palsson, B. Ø. 2007 A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121.

50. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. 2015 BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* , gkv1049–.

51. Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. 2013 Sybil--efficient constraint-based modelling in R. *BMC Syst. Biol.* 7, 125.

52. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. 2014 RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* 42, D553–9.

53. Trudeau, D. L., Smith, M. A. & Arnold, F. H. 2013 Innovation by homologous recombination. *Curr. Opin. Chem. Biol.* 17, 902–9.

54. Apic, G. & Russell, R. B. 2010 Domain recombination: a workhorse for evolutionary innovation. *Sci. Signal.* 3, pe30.

55. Wagner, A. 2012 The role of robustness in phenotypic adaptation and innovation. *Proc. R. Soc. B Biol. Sci.* 279, 1249–1258.

56. Fares, M. A. 2015 Survival and innovation: The role of mutational robustness in evolution. *Biochimie* 119, 254–61.

57. Wagner, A. 2011 The molecular origins of evolutionary innovations. *Trends Genet.* 27, 397–410.

58. Jacob, F. & Monod, J. 1961 Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3, 318–56.

59. Itoh, T., Takemoto, K., Mori, H. & Gojobori, T. 1999 Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* 16, 332–46.

60. Price, M. N., Arkin, A. P. & Alm, E. J. 2006 The life-cycle of operons. *PLoS Genet.* 2, e96.

61. Hunter, R. L. & Market, C. L. 1957 Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science* 125, 1294–5.

62. Khersonsky, O. & Tawfik, D. S. 2010 Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* 79, 471–505.

63. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. 2010 Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol. Syst. Biol.* 6, 436.

64. Nam, H., Lewis, N. E., Lerman, J. A., Lee, D.-H., Chang, R. L., Kim, D. & Palsson, B. O. 2012 Network context and selection in the evolution to enzyme specificity. *Science* 337, 1101–4.

65. Fang, G., Rocha, E. P. C. & Danchin, A. 2008 Persistence drives gene clustering in bacterial genomes. *BMC Genomics* 9, 4.

66. Fang, G., Rocha, E. & Danchin, A. 2005 How essential are nonessential genes? *Mol. Biol. Evol.* 22, 2147–56.

## 2.7. Supplementary Information

### Text S1: Supplementary Methods

#### *(a) Genome-scale metabolic networks and their phenotypic representations*

The set of genomically encoded biochemical reactions proceeding inside a given organism constitutes an organism's metabolic genotype [1–3]. This genotype enables an organism to extract energy and produce small biomass building blocks, such as amino acids, from extracellular nutrients. Reconstruction of this genotype from genomic and biochemical information has been successful for multiple organisms [4–7].

Each metabolic network contains a subset of the "reaction universe" of all biochemical reactions that take place in the biosphere (See Text S1b). We have curated a representation of this universe, which comprises 5906 reactions and is based on current metabolic knowledge [8–12]. We represent an organism's metabolic genotype as a binary vector of length 5906. Each entry of this vector corresponds to a given reaction in the reaction universe, and is equal to one if the corresponding reaction is present in the metabolic network, and zero otherwise. Thus, each genotype can be thought of as a single member of a vast space of all possible metabolic networks, which contains $2^{5906}$ distinct genotypes. We define the phenotype of a given metabolic genotype based on its viability on 50 distinct minimal environments that differ only in the carbon source (See Text S1c). We consider that a genotype is *viable* on a given carbon source, if it can produce all the essential biomass precursors from the given carbon source, and we use Flux Balance Analysis (FBA, See Text S1d) to determine viability [12–14]. We represent the phenotype of a given metabolic genotype as a binary vector of length 50. Each entry of this vector corresponds to a given carbon source, and it is equal to one if the genotype is viable on this carbon source, and zero otherwise.

***(b) Reaction universe***

The reaction universe is a set of metabolic reactions known to occur in some organism. For the construction of this universe, we used data from the LIGAND database [9,15]of the Kyoto Encyclopedia of Genes and Genomes [11,16]. Briefly, the LIGAND database, which is comprised of the REACTION and the COMPOUND databases, provides information on reactions, associated stoichiometric information, chemical compounds involved, and the Enzyme Classification (E.C.) identifier of each reaction. We used the REACTION and the COMPOUND databases to construct our universe of reactions, and excluded (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and, (iv) reactions involving complex metabolites without chemical information about their structure [8]. The published *E. coli* metabolic model (iAF1260) consists of 1397 non-transport reactions [12]. We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the unique (non-duplicate) reactions. This resulted in a universe of reactions consisting of 682 transport, 5906 non-transport reactions and 5030 metabolites.

***(c) Chemical environments***

We consider 50 minimal growth environments, each of which included oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe2+ and Fe3+), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese, zinc, and a specific carbon source Importantly, to represent different chemical environments, we vary the carbon source while keeping all other nutrients constant. We consider a metabolic network viable on a given carbon source, if it can synthesize all essential biochemical precursors when this carbon source is provided as the sole carbon source in a minimal medium.

We used 50 carbon sources for our analysis of randomly sampled metabolic networks, including the following 27 glycolytic carbon sources: D-Glucose, D-Glucose 6-phosphate, Trehalose, Maltose, Lactose, D-Fructose 6-phosphate, D-Fructose, D-Mannose, D-Mannitol, D-Glucose 1-phosphate, D-Sorbitol, Maltotriose, D-Allose, D-Ribose, D-Xylose, D-Gluconate, 5-dehydro-D-Gluconate, L-Rhamnose,

L-Fucose, L-Arabinose, L-Lyxose, D-Galactose, Melibiose, D-Galactonate, N-Acetyl-D-glucosamine, N-Acetyl-D-mannosamine, N-Acetylneuraminate.

In addition, we used the following 23 gluconeogenic carbon sources: Pyruvate, L-Alanine, L-Lactate, D-Alanine, D-Malate, Acetate, L-Serine, L-Malate, D-Serine, Glycine, Glycolate, L-Aspartate, Succinate, Fumarate, 2-Oxoglutarate, D-Galacturonate, D-Galactarate, D-Glucarate, L-Galactonate, D-Glucoronate.

And we used 3 nucleoside carbon sources: Adenosine, Deoxyadenosine, Inosine.

For the analysis of prokaryotic metabolic networks in the BiGG database, we used the following 137 carbon sources:

Acetaldehyde, Acetate, Acetoacetate, Adenine, Adenosine, Allantoin, Bicarbonate, Biotin, Butyrate (n-C4:0), Carbonic acid, Choline, Citrate, Cyanate, Cytidine, Cytosine, D-Alanine, D-Fructose, D-Galactarate, D-Galactonate, D-Galactose, D-Galacturonate, D-Glucarate, D-Gluconate, D-Glucosamine, D-Glucose, D-Glucose 6-phosphate, D-Glucuronate, D-Glyceraldehyde, D-Lactate, D-Mannitol, D-Mannose, D-Mannose 6-phosphate, D-Methionine, D-Ribose, D-Serine, D-Sorbitol, D-Xylose, Deoxyadenosine, Deoxycytidine, Deoxyguanosine, Deoxyinosine, Deoxyuridine, Dihydroxyacetone, Dimethyl sulfide, Dimethyl sulfoxide, Ethanol, Folate, Formate, Fumarate, Galactitol, Gamma-butyrobetaine, Glycerol, Glycerol 3-phosphate, Glycine, Glycine betaine, Glycolate, Guanine, Guanosine, Hexadecanoate (n-C16:0), Hypoxanthine, Indole, Inosine, L-Alanine, L-Arabinose, L-Arginine, L-Asparagine, L-Aspartate, L-Carnitine, L-Cysteine, L-Fucose, L-Fucose 1-phosphate , L-Glutamate, L-Glutamine, L-Histidine, L-Idonate, L-Isoleucine, L-Lactate, L-Leucine, L-Lysine, L-Malate, L-Methionine, L-Phenylalanine, L-Proline, L-Rhamnose, L-Serine, L-Threonine, L-Tryptophan, L-Tyrosine, L-Valine, L-tartrate, Lactose, Maltohexaose, Maltopentaose, Maltose, Maltotetraose, Maltotriose, Melibiose, Meso-2,6 Diaminoheptanedioate, Methanol, N-Acetyl-D-glucosamine, N-Acetyl-D-mannosamine, N, Acetylneuraminate NMN, Nicotinamide adenine dinucleotide, Octadecanoate (n-C18:0), Ornithine, Phenylpropanoate, Pimelate, Protoheme, Putrescine, Pyruvate, Riboflavin, Spermidine, Succinate, Sucrose, Taurine, Tetradecanoate (n-C14:0),Thiamin, Thymidine,Trehalose, Trimethylamine, Trimethylamine N-oxide, Uracil, Urea, Uridine, Xanthine, Xanthosine, AMP, (R)-Pantothenate, S)-Propane-1,2-diol, 1,5-Diaminopentane, 2-Dehydro-3-deoxy-D-

gluconate, 2-Oxoglutarate, 3-(3-hydroxy-phenyl)propionate, 3-hydroxycinnamic acid.

***(d) Flux balance analysis***

Flux balance analysis (FBA) is a computational method that is widely used for the quantitative analysis and modeling of metabolic networks [13]. Based on the stoichiometric coefficients of the metabolites participating in the reactions of a given metabolic network, FBA predicts the metabolic flux through each reaction. Stoichiometric coefficients are stored in a stoichiometric matrix *S*, which is of dimension $m \times n$, where $m$ and $n$, respectively, denote the number of metabolites and the number of reactions in a metabolic network. FBA constrains the flux through each reaction based on the assumption that a metabolic network is in a steady state where metabolite concentrations do not change, i.e., $Sv = 0$, where $v$ is the vector of metabolic fluxes $v_i$ through reaction $i$. The solutions of the equation $Sv = 0$, that is, the nullspace of matrix *S*, comprises all flux vectors that are allowable in steady state. The null space is further constrained by physicochemical information regarding the maximum and minimum possible flux through each reaction. FBA relies on an optimization procedure called linear programming to identify those flux vector(s) among the allowable ones that maximize an objective function *Z*. This task can be formulated as finding a flux vector $v^*$ with the property

$$v^* = max_v\, Z(v) = max_v \{ c^T v \mid Sv = 0,\ a \leq v \leq b \},$$

where the vector $c$ contains a set of scalar coefficients representing the maximization criterion, and each entry $a_i$ and $b_i$ of vectors $a$ and $b$, respectively, indicates the minimally and maximally possible flux through reaction $i$. The vector $c$ represents the proportions of each small biomass molecule in a cell's biomass. Therefore $v^*$ maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass [14]. Here we use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and we consider a metabolic network viable if it can produce all essential biomass precursors. In a free-living bacterium like *E. coli,* there are approximately 60 such molecules including 20 amino acids, DNA, and RNA precursors, lipids and cofactors. We used the biomass composition of the *E. coli* metabolic model iAF1260 to define the vector $c$ [12]. Moreover, we used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and

CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problem of FBA.

The major limitation of FBA is that it neglects regulatory constraints that can arise through suboptimal expression or regulation of enzymes. Newly horizontally transferred genes cannot easily establish regulatory interactions with their host genes, and it may thus take considerable adaptive evolution until they become expressed at a maximal or optimal level [17]. Such regulatory constraints would be especially important if we focused on quantitative predictions of biomass growth [18]. However, we use FBA solely for qualitative prediction of viability. This focus on qualitative phenotypes is biologically sensible. The reason is that many organisms grow slowly in their native environment [19–21], implying that regulation for maximal biomass production is far from universal. Moreover, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments [18,22,23].

***(e) Generation of random metabolic networks***

We here employ a previously described *in silico* process which relies on Markov Chain Monte Carlo (MCMC) random walks to generate metabolic networks that comprise random sets of metabolic reactions that are viable on a given carbon source [8,24]. This procedure can produce metabolic networks that are sampled uniformly from the set of all metabolic networks viable on a given carbon source [8,24]. Briefly, in each step of such a random walk we perform a reaction swap, which is defined as altering a metabolic network by adding a randomly chosen reaction from the reaction universe, and then deleting a reaction randomly chosen from the set of reactions present in the metabolic network. If the reaction swap disrupts the metabolic network's viability on the given carbon source (as determined by FBA) we reject it, and perform another reaction swapping until we find a reaction swap that does not disrupt viability. This procedure also ensures that the total number of reactions remains constant. For the MCMC method to produce random samples of metabolic networks, it is essential to carry out enough reaction swaps to "erase" the random walker's similarity to the initial metabolic network. Previously, it has been shown that $3 \times 10^3$ reaction swaps are sufficient for this purpose [8,24]. Each of our random walks starts from *E. coli*'s metabolic network and performs $10^4$ reaction swaps before

storing the final metabolic network for further analysis. We used $10^4$ independent random walks conducted in this way to create $10^4$ random metabolic networks viable on glucose. We used the same procedure to generate $10^4$ random metabolic networks viable on acetate.

***(f) Generation of parental metabolic network pairs***

Our analyses required us to recombine pairs of "parental" metabolic networks with particular features, such as (i) their genotypic distance ($D$), defined as the number of reactions differing between the parents, (ii) their phenotypic complexity ($||P||$), that is, the number of carbon sources on which they are viable, (iii) their phenotypic distance ($\Delta P$), that is, the number of carbon sources on which only one but not the other member of a parental pair is viable, and (iv) their genotypic complexity ($||G||$, or metabolic network size), defined as the number of reactions in each metabolic network pair.

To identify parental metabolic networks with a given $\Delta P$ and $||P||$ we first selected, among all $\binom{10^4}{2}$ possible random metabolic network pairs that can be formed from $10^4$ MCMC-sampled metabolic networks, those pairs that are viable on exactly $||P||$ carbon sources and that have a given $\Delta P$. We then randomly chose from them a set of 1000 pairs for further analysis.

Less straightforward than identifying parental metabolic networks with a given $\Delta P$ and $||P||$ is to identify those with a given genotypic distance ($D$), because the random metabolic networks generated by MCMC sampling generally have genotypic distances sufficiently large ($D \approx 2000$) to be biologically unrealistic for modeling frequently recombining prokaryotic genomes. To create less diverse metabolic network pairs, we took two different MCMC random walk approaches that yielded similar results. The first revolves around a reaction-swapping random walk starting with a pair of randomly chosen metabolic networks from our sample of $10^4$ sampled metabolic networks. In each step of this random walk, we subjected each parental metabolic network to a reaction swap, and we accepted each reaction swap if it (i) preserved the original phenotype, and (ii) did not increase the genotypic distance of the two metabolic networks after the swap, otherwise we rejected the reaction swap. We continued this procedure until the genotypic distance between the metabolic

networks became equal to a desired distance $D$. This approach is very time-consuming. The second approach is much faster and uses a more biologically inspired mechanism to generate metabolic networks (see Text S2 [24,25]), but it also suffers from a technical limitation (Text S2), which is why we report mostly on the first approach.

Finally, to generate parental metabolic networks with a given number of reactions $||G||$ we started from a random viable metabolic network generated by MCMC sampling, as described in the Text S1e. All such metabolic networks have the same number of reactions as *E. coli* (2079). We then applied a sequence of reaction deletions that preserved viability on glucose (or acetate, depending on analysis) until we reached the desired $||G||$. Then, we sampled pairs of metabolic networks with a given $D$, $\Delta P$ and $||P||$ among the metabolic networks with $||G||$ reactions in the manner described above.

***(g) Modeling recombination and mutation in metabolic networks***

Prokaryotic genomes undergo recombination via horizontal gene transfer [26], whose incidence is large and greater than that of point mutations [27–29]. It changes the organization and gene content of genomes on short evolutionary time scales [26,30,31], and can involve very distantly related organisms [32,33]. Various mechanisms of horizontal gene transfer add genes unidirectionally from a donor to a recipient, but incorporating such genes into the recipient genome relies on recombination [26]. The genomes of many prokaryotes frequently undergo homologous recombination, that is, a reciprocal exchange of DNA segments between DNA sequences [34]. Because such recombination can also delete genes, and because of a general deletion bias in prokaryotic genomes [35], prokaryotic recombination involves gene loss as well as gene gain. What is more, the majority of newly acquired genes obtained via horizontal gene transfer reside in the genome only for short amounts of time [36]. Motivated by these observations, we here model prokaryotic recombination as a process where the transfer of reactions from a donor to a recipient metabolic network is compensated by deletion of other reactions from the recipient.

To model recombination for each parental metabolic network pair, we generated 1000 recombinant offspring by (i) adding to the recipient metabolic network a given number $n/2$ of randomly chosen reactions that were present in the donor and absent in

the recipient, followed by (ii) deleting *n/2* reactions randomly chosen from the recipient. Thus, the total number of reactions changed by a recombination event in the recipient is equal to *n*. For reasons of computational feasibility, we analyzed only recombinant pairs where the probability that a recombination event preserves viability exceeded $10^{-3}$. Text S3 and figure S1 show that this is the case for values of *n* up to *60,* which is why we chose *n=60* as the highest amount of reaction changes during a recombination event. Empirical observations also suggest that this number of reactions would not be unrealistically large, because horizontal gene transfer can affect long DNA regions[44]. Transferred material that is integrated into the host genome by recombination can constitute stretches of non-coding DNA, fragments of genes [37,38], entire genes [39], multiple adjacent genes [40,41], operons, transposable chromosomal elements, plasmids, as well as other naturally occurring extrachromosomal elements [42]. The length of contiguous transferred stretches may range from a few nucleotides [43] to more than 3 Mbp [44], i.e., some two thirds of the length of the *E. coli* genome, which encodes more than 1300 reactions. In addition, some megabase-scale horizontally transferred genes can become incorporated into a chromosome in the form of hundreds of smaller fragments [45].

To implement an amount of random mutational metabolic change that is comparable to the same amount of recombinational change, for a given number of altered reactions (*n*) we created a "mutational" offspring of a metabolic network by adding *n*/2 randomly chosen reactions from the reaction universe, and deleting *n/2* randomly chosen reactions among the set of reactions present in the metabolic network. Note the key difference between mutation and recombination: In recombination the *n/2* reactions that are added to the recombinant offspring are chosen randomly from another viable metabolic network (the donor), whereas in mutation they are taken from the whole reaction universe.

### *(h) Genomic recombination in prokaryotic metabolic networks from the BiGG database*

We validated our observations based on randomly sampled viable metabolic networks by considering the genome-scale metabolic networks of 61 bacterial species available at the BiGG database [46], using the R-package Sybil [47]. For this analysis, we generated a reduced universe of reactions comprised of the union of the sets of

reactions present in the 61 metabolic networks. This universe altogether contains 3404 internal reactions, 3156 transport reactions, and a different biomass reaction for each organism. As potential carbon sources, we used all 137 carbon-containing metabolites that occurred as metabolites external to at least one organism in the database, and thus assigned a phenotype vector of length 137 to each metabolic network using FBA.

To model recombination among the metabolic networks of these 61 organisms, we used one main approach, which incorporates information about the linkage of the genes encoding metabolic reactions. To this end, we used the gene-reaction association rules defined in the BiGG database for each organism (in MAT files, grRules) [46], and ordered the genes in each organism based on their genomic position, as obtained from the RefSeq microbial genome database [48].

For a specific recombination event between a donor and a recipient organism, we first chose at random a stretch of DNA from the donor organism that contains a given number of metabolic genes. To generate a recombinant offspring we added this stretch of DNA to the recipient, and subsequently deleted a randomly selected stretch of DNA from the recipient genome. We translated the added and deleted genes into reactions based on the gene-reaction rules for the donor and recipient organism. We set the number of genes in every donor DNA stretch such that on average (among all recombination events between all metabolic network pairs) a given number of $n$ reactions are added to the recipient metabolic network, and an equal number $n$ of reactions are deleted from it. Because gene-reaction associations are not generally one-to-one and can be very complicated, and because most of the reactions that are encoded in a given stretch of DNA may already be present in the recipient metabolic network, the number of metabolic genes in donor DNA required for adding $n$ reactions will be higher than $n$ (usually $\approx 2n$). In contrast, we found that including $\approx 0.9n$ metabolic genes into a DNA stretch to be deleted from the recipient genome usually sufficed to eliminate $n$ reactions from the recipient metabolic network, because deletion of a single metabolic gene often causes elimination of multiple reactions

In a second approach for recombining prokaryotic genomes, we neglected linkage between metabolic genes and added or deleted reactions randomly, just as we had

done for randomly sampled viable metabolic networks, irrespective of the genomic position of metabolic genes encoding these reactions.

**Text S2: An alternative MCMC approach to generate parental metabolic networks with a given genotypic distance (*D*)**

In addition to our first and main approach (see text S1f) for creating metabolic networks with a given genotypic distance D, we also pursued a second approach. This second approach starts from a parental metabolic network $M_1$ and generates a recombination partner $M_2$ through a sequence of MCMC random walks, which preserves the phenotype of $M_1$ but increases the genotypic distance to $M_2$, until a desired $D$ between $M_1$ and $M_2$ is reached. Because this method resembles the divergence of species from a common ancestor, it is biologically motivated, but it has a technical limitation that is associated with inactive (blocked) reactions – reactions having zero metabolic flux for stoichiometric reasons [24,25]. Probably due to the shorter MCMC random walks in this second approach, a greater percentage of the $D$ reactions that are not shared between two metabolic networks are blocked reactions in the second approach (90.09%) compared to the first one (66.78% of reactions). The innovation potential of inactive reactions is almost negligible in comparison to active reactions (Figure S2), and the fraction of innovative offspring is thus considerably (almost an order of magnitude) lower for the second approach than for the first approach.

To render results from the two approaches comparable, one can adjust the ratio of inactive reactions in the $D$ non-shared reactions between metabolic network pairs, as illustrated in the following example. Let us assume a given parental metabolic network pair obtained with the second method has $D=1000$ non-shared reactions, and since ≈90% of these reactions are inactive (blocked), only around 100 of them will be active. To make the ratio of active to inactive reactions among these non-shared reactions equal to the 66.78% (≈2/3) that are characteristic of parental metabolic network pairs obtained with the first method, one would require almost 200 inactive reactions (200/(100+200)= 2/3). Thus, whenever one wants to transfer a given number of reactions ($n$) from donor to recipient, one can first select 200 inactive reactions from the 900 inactive reactions and then select the $n$ reactions to be recombined from a set that includes these 200 inactive reaction and 100 active

reactions. This ensures that a comparable proportion of active reactions are transferred from donor to recipient in the two approaches. After this adjustment, the two approaches yield virtually identical observations (Compare figure1 with figure S3). However, the manipulations required in the second approach make it less useful. We therefore chose to rely on the first approach throughout this study.

**Text S3: Robustness of genome-scale metabolic networks decreases exponentially with increasing the number of deleted reactions**

Recombination that involves both the addition and deletion of reactions has the potential to create inviable recombinant offspring. The greater the number of reactions that are deleted in a recombination event, the greater will be this fraction of inviable offspring. Before embarking on a systematic analysis of recombination's effects, we needed to find out how large the number of reactions deleted in a recombination event (*n/2*) can become, before the number of viable offspring becomes too small for computational analysis. To this end, we generated 1000 random genome-scale metabolic networks that we required to be viable only on glucose as the sole carbon source. For each of these randomly sampled viable metabolic networks and for each value of *n* between one and sixty, we created 1000 offspring in which we deleted *n/2* randomly chosen reactions. Figure S1 shows a box plot of the fraction of metabolic networks that remain viable on glucose as the sole carbon source after this procedure.

The fraction of viable metabolic networks declines exponentially with the number of deleted reactions. For $(n/2)>30$ the fraction of metabolic networks that retain viability becomes very low, e.g., it declines below 0.001 for viability on glucose, such that fewer than one of 1000 offspring would be viable on glucose. At numbers beyond $(n/2)>30$, the number of recombination events needed to create any viable metabolic networks becomes computationally prohibitive. For this reason, we chose $n=60$ as the highest value of *n* for our recombination analysis.

**Text S4: The rate of recombination between bacterial species decreases exponentially by increasing metabolic distance**

We wanted to obtain a crude estimate of the relationship between the metabolic distance of two bacterial species and the likelihood that such species undergo a

successful homologous recombination event. To estimate this relationship, we pursued a three-step procedure.

In the first step, we estimated the DNA-based genotypic distance between two bacterial species whose metabolic networks differ by a given number of reactions. To this end, we used curated metabolic networks from 51 bacterial species, which had been obtained through state-of-the art techniques for genome annotation, generation of biomass reactions, reaction network assembly, and thermodynamic analysis of reaction reversibility [49]. We define the normalized metabolic genotype distance *d* of two prokaryotes as the number *D* of reactions differing between their metabolic networks, divided by the total number of reactions present in at least one of the two metabolic networks and computed this distance for all pairs of the 51 metabolic networks. On average, a relative metabolic distance of $d=0.1$ corresponds to an absolute difference of $D\cong150$ in reaction number, but we note that the relationship between *d* and *D* depends on the total number of reactions in each metabolic network.

We then aimed to relate metabolic divergence to DNA sequence divergence between these species. To this end, we used the housekeeping gene rpoB, which encodes the β-subunit of RNA polymerase. We obtained the rpoB coding sequences for these 51 species from NCBI (http://www.ncbi.nlm.nih.gov), and aligned them with the PAL2NAL web server, which provides robust alignment of DNA sequences based on the corresponding protein sequences [50]. We then computed all pairwise Hamming distances from the aligned rpoB sequence alignment for these 51 species, normalized these quantities to the interval (0,1), and used them as our measure of sequence divergence.

We note that even species with modest sequence divergence can have considerable metabolic distance. For example, the species pair *Buchnera aphidicola* and *Yersinia pestis* have an *rpoB* DNA sequence distance of 0.29, but a metabolic distance *d* of 0.65, which corresponds to an absolute difference of 1004 reactions. Examples of moderately high metabolic divergence exist even from strains of the same organisms, such as *Streptococcus pneumoniae* TIGR4 and R6, which differ in 64 reactions or a fraction $d=0.068$ of their metabolic networks. At greater sequence distances of 0.45, metabolic distances reach values up to $d=0.69$ (e.g., *Yersinia pestis* and *Rickettsia*

*prowazekki* have *rpoB* DNA sequence divergence 0.45 and a metabolic distance of *d=0.684*, corresponding to 1066 reaction differences.) Metabolic distance and sequence divergence are significantly correlated (Pearson's $r=0.60$, $P<10^{-40}$), and a linear regression analysis (red line in figure S4a) yields a regression coefficient of 0.82 with an intercept of 0.1. We use this regression analysis to translate metabolic distance into sequence divergence and vice versa.

In the second step, we took advantage of experimental data on the exponential relationship between the likelihood of a successful recombination event and (rpoB-based) sequence divergence between recombining species [33,51]. Specifically, we used such data for 19 species pairs in the genera Bacillus and Streptococcus. Figure S4b shows that the logarithm of the relative recombination rate decreases linearly with increasing sequence divergence between the donor and recipient species. A linear regression analysis (black line in the figure) yields a regression coefficient of -18.40 with an intercept of 0.11.

In the third step, we integrated data from step one and two to relate metabolic distance to the likelihood of a successful recombination event (Figure S4c). The figure shows that the logarithm of the relative recombination rate linearly decreases with increasing metabolic distance between the donor and recipient species. A linear regression analysis (red line in the figure) yields a regression coefficient of -22.57 with an intercept of 0.62. In sum, sequence and recombination data suggests that the likelihood of a successful recombination event between two species would decrease exponentially with their metabolic distance. This also holds if we exclude endosymbiotic or host-associated pathogens from our analysis.

Importantly, we note that metabolic distance will not be the only determinant of successful recombination between bacteria of different species. Part of the reason is that only a minority of genes in any bacterial genome are typically involved in metabolic network (e.g., 31% in *E. coli*). In addition, other incompatibilities, such as those between restriction-methylation systems [52] or DNA repair mechanisms [53] may hinder recombination. Our analysis merely goes to show that the minimal recombination distances of *D=100* we use are not unrealistically low. Many bacteria that would successfully recombine in the wild have greater metabolic distances (Figure S4c).

**Text S5: Phenotypically more diverse parental metabolic networks are more likely to generate metabolically innovative offspring**

We asked whether the phenotypic diversity of recombining parents influences the incidence of innovative offspring. On the one hand, recombining parents viable on the same combination of carbon sources might create a greater fraction of viable offspring, which might also increase the incidence of offspring with novel metabolic abilities. On the other hand, recombining parents viable on different combinations of carbon sources might produce recombinant offspring with a greater number of *novel* reaction combinations, and thus a greater number of metabolic innovations.

To find out whether one of these hypotheses is correct, we created pairs of metabolic networks at a fixed genotypic distance ($D=100$), but with different metabolic phenotypes $P_1$ and $P_2$ and with identical phenotypic distances ($\Delta P$), that is, identical number of carbon sources on which one parent is viable but the other isn't, or vice versa. To prevent confounding our analysis by the number of carbon sources $||P||$ on which a metabolic network is viable, we kept $||P||$ constant and required that each parent was viable on exactly 10 carbon sources. (In other words, all metabolic networks in this analysis are viable on glucose and on nine other carbon sources.) We then varied $\Delta P$ in four steps between 0 and 16, created 1000 metabolic network pairs for each value of $\Delta P$, and from each pair we created 1000 recombinant offspring in which $n$ reactions were altered through recombination. We then determined for each offspring whether it was viable on any carbon source that neither of the parents were viable on. Figure 2c (main text) shows that regardless of the number $n$ of altered reactions, the fraction of innovative offspring ($f_{innov}$) increases with the phenotypic distance $\Delta P$ among parents.

The increase of innovation with parental phenotypic diversity cannot just be explained by a greater fraction of viable offspring, because parental phenotypic diversity does not influence this fraction (Figure S10a and S10b). In contrast, as $\Delta P$ increases, so does the fraction of reactions with $I_{SE} > 0.5$ that can potentially be transferred from donor to recipient (Figure 2d (main text)), once again highlighting the role of this process in innovation. Parental phenotypic diversity $\Delta P$ does have no impact on the number of carbon sources on which innovative offspring gain viability.

Specifically, we observed that innovative offspring typically gains viability on two to three additional carbon sources, and shows an average phenotypic distance between four and five, regardless of whether it arose through recombination or mutation, and independent of parental genotypic or phenotypic features.

We complemented these analyses by focusing on an alternative way of defining phenotypic heterogeneity that is based on viability on two specific classes of carbon sources, namely those involved primarily in glycolysis, and those involved primarily in gluconeogenesis (See Text S1c). We found that offspring of parents viable on different classes of carbon sources display a greater incidence of innovation, compared to offspring of parents that are viable on the same class of carbon sources (Text S6).

**Text S6: Parental metabolic networks viable on different classes of carbon sources are more likely to generate innovative offspring than parents that are viable on the same classes of carbon sources**

In this analysis, we focused on two specific classes of carbon sources, namely those involved primarily in glycolysis, and those involved primarily in gluconeogenesis (Text S1c). In a previous contribution, we had shown that metabolic networks required to be viable on one glycolytic (gluconeogenic) carbon source tended to be viable also on other glycolytic (gluconeogenic) carbon sources [54]. We wanted to find out whether parental viability on either glycolytic, gluconeogenesis, or both kinds of carbon sources influenced the incidence of novel metabolic traits in the offspring. To this end, we created 1000 pairs of donor – recipient metabolic networks (genotype distance $D=100$) with each of the following properties (i) both parents are viable on five glycolytic carbon sources, (ii) both parents are viable on five gluconeogenic carbon sources, (iii) all donor metabolic networks are viable on five gluconeogenic carbon sources, and all recipient metabolic networks are viable on five glycolytic carbon sources, and (iv) all donor metabolic networks are viable on five glycolytic carbon sources, and all recipient metabolic networks are viable on five gluconeogenic carbon sources. To exclude parental phenotypic diversity as a confounding factor, we ensured that it had a constant value of $\Delta P=10$ for all parents

in all three categories. Aside from these constraints, we chose glycolytic and gluconeogenic carbon sources at random.

For each pair of metabolic networks we created 1000 offspring with a fixed number of altered reactions, and found that recombinants of parents viable on different kinds of carbon sources (i.e. gluconeogenic-glycolytic) display a greater incidence of innovation (Figure S11a). This greater incidence of innovation cannot solely be explained by a greater fraction of viable offspring, because parental viability on different classes of carbon sources does not influence the fraction of viable offspring (Figure S11b). Thus, we conclude that phenotypically more heterogeneous parental metabolic networks are more likely to generate innovative recombinants.

**Text S7: Phenotypically less complex parental metabolic networks are more likely to generate metabolically innovative offspring**

We also investigated the impact of phenotypic complexity on metabolic innovation. We define the complexity of a phenotype $P$ as the number $||P||$ of carbon sources on which it is viable. For this analysis, we generated parental metabolic networks with the same genotypic distance $D=100$ but with varying phenotypic complexity. In addition, we required that both metabolic networks in a pair are viable not only on the same number of carbon sources, but also on the exact same carbon sources. Specifically, we analyzed 1000 pairs of random parental metabolic networks viable on $||P||=1$, 5, or 10 carbon sources. For each of the 1000 pairs at each value of $||P||$, we created 1000 recombinant offspring with $n$ altered reactions. Figure 2e (main text) shows that the fraction of innovative offspring ($f_{innov}$) decreases with increasing phenotypic complexity. The more carbon sources a metabolic network is viable on, the smaller the likelihood that recombination creates viability on further carbon sources. This difference is not simply caused by a decrease in the fraction of viable offspring with increasing phenotypic complexity (Figures S12a and S12b). Also, $||P||$ does not impact the number of additional carbon sources that an innovative offspring gains viability on. The fraction of exchangeable reactions with $I_{SE} > 0.5$ ($f_{super}$) decreases with increasing phenotypic complexity (Figure 2f (main text)).

We also wished to analyze the effect of phenotypic complexity on metabolic innovation for metabolic networks viable on more than 10 carbon sources (i.e.,

$||P||=20$, 30, 40). However, creating 1000 metabolic network pairs with these values of $||P||$ that were viable on the exact same combination of carbon sources was computationally infeasible. We thus created 1000 metabolic network pairs whose phenotype vectors differed in a fixed number of 10 non-zero entries. For example, in such a pair with $||P||=30$, both members would be viable on the same 25 carbon sources. In addition one member would be viable on five carbon sources that the other one is not viable on, and vice versa. Furthermore, we required a fixed genotypic distance $D=100$ for all metabolic network pairs in this analysis. For each value of $||P||$ subject to these constraints, we created 1000 recombinant offspring with $n$ altered reactions for each of the 1000 parental metabolic network pairs. Consistent with data from figure 2e (main text), the fraction of innovative offspring ($f_{innov}$) decreases with increasing phenotypic complexity (Figure S13a). Figures S13b and S13c show that this difference is not simply caused by a decrease in the fraction of viable offspring with increasing phenotypic complexity. Figure S13d shows that fraction of reactions with superessentiality higher than 0.5 ($f_{super}$) decreases with increasing phenotypic complexity.

**Text S8: Larger parental metabolic networks are recombinationally more robust and so more likely to generate metabolically innovative offspring**

We define the genotypic complexity of a metabolic network as the number of reactions ($||G||$) present in this metabolic network. We wanted to find out whether it affects recombinational robustness and the incidence of novel phenotypes among recombinant offspring. To this end, we analyzed metabolic networks with sizes that vary between $||G||=1500$ and $||G||=2000$ reactions. Specifically, we created 1000 random viable donor recipient pairs with constant genotype distance $D=100$ for each size class, where we required the parental metabolic networks to be viable only on glucose. We then created from each parental pair 1000 recombinant offspring with a specific number $n$ of altered reactions.

Figure S14a shows that recombinational robustness increases with increasing genotypic complexity of the parents. For example, at $n=10$ recombined reactions the fraction of viable offspring is three times higher for parental metabolic networks with $||G||=2000$ reactions than for parental metabolic networks with $||G||=1500$ reactions (0.3 vs. 0.1, Figure S14a). We observe the same result, when we use parental

metabolic networks viable on acetate for this analysis (Figure S14b). Moreover, we observed that the fraction of innovative offspring ($f_{innov}$) also increases with increasing metabolic network size $||G||$ (Figure S14c). Again this result does not change if parental metabolic networks are viable on acetate instead of on glucose (Figure S14d).

In sum, unlike other quantities such as genotypic distance and phenotypic diversity of parents, which do not impact recombinational robustness, and thus influence innovation directly, parental genotypic complexity ($||G||$) increases recombinational robustness, and can thus enhance innovation indirectly by increasing robustness.

**Text S9: Effects of genotypic and phenotypic features of prokaryotic parental metabolic networks on recombinational robustness and innovation**

Unlike our analyses of random viable metabolic networks, where we were able to control genotypic parameters such as the number of reactions, and phenotypic parameters such as phenotypic complexity by sampling genotype space appropriately, these parameters are fixed properties of the 61 specific prokaryotic metabolic networks we analyzed. Moreover, when analyzing random viable networks we could sample metabolic network pairs that varied in only one parameter, which is not possible for prokaryotic metabolic networks. However, to control relevant parameters to some extent, we took the following steps.

First, to prevent size variation in metabolic networks from confounding our analysis, we observed that the majority (47 of 61) of prokaryotic metabolic networks show a narrow size range between 1250 and 1350 internal reactions (Figure S16a), and focused our analysis on these metabolic networks. (278 among all $\binom{47}{2}$ possible pairs of these metabolic network pairs have at least one offspring that is viable on a new carbon source.)

Second, we observed that the distribution of parental genotypic distance ($D$), phenotypic distance ($\Delta P$), and phenotypic complexity ($||P||$) is distinctly bimodal for these 278 parental metabolic networks (Figures S16b, S16c, and S16d). No parental metabolic network has intermediate values for any of these parameters, such that metabolic networks can be subdivided into “high” and ”low” categories for each

parameter. Moreover, metabolic networks with high $\Delta P$ have low $||P||$ (Figure S16e). Based on these observations, we subdivided parental metabolic network pairs into 4 categories: i) high $D$ and low $\Delta P$ (high $||P||$), ii) low $D$ and low $\Delta P$ (high $||P||$), iii) high $D$ and high $\Delta P$ (low $||P||$), iv) low $D$ and high $\Delta P$ (low $||P||$). The number of parental metabolic networks in categories (i) through (iv) was 96, 106, 12, and 64, respectively. Recombinational robustness differs little among metabolic networks in these four categories (Figures S17b and S17c), and so these parameters do not strongly influence robustness, which is consistent with our observations from random viable metabolic networks. In contrast, the fraction of innovative offspring of parental metabolic networks in the fourth category (with low $D$ and high $\Delta P$ (low $||P||$)) is highest, and it is lowest for metabolic networks in the first category (with high $D$ and low $\Delta P$ (high $||P||$)) (Figures 3c and S17a). This is again consistent with our observations from random viable metabolic networks, where parents with low genotypic distance, high phenotypic distance, and low phenotypic complexity are more likely to generate innovative offspring.

**Text S10: Superessential reactions can explain the effect of parental genotypic and phenotypic diversity and complexity on metabolic innovation.**

Superessential reactions that are involved in recombination events can explain a series of patterns in our data. The first is that a given number of reaction changes can elicit more metabolic innovation when caused by recombination rather than by mutation. While recombination adds reactions to a recipient that already occur in a (viable) donor, random mutations add reactions unrelated to the metabolic network of the donor. Because this reaction universe contains fewer highly super-essential reactions than any donor, adding reactions from it is less likely to yield innovations (electronic supplementary material, figure S18). Put differently, when recombination introduces new metabolic reactions into an organism, it preferentially introduces reactions that have been "pretested" by evolution, because they form part of a related viable genotype. In contrast, mutations may introduce reactions that are incompatible with this genotypic background, in the sense that they cannot interact productively with it. This observation is consistent with observations from other systems, such as proteins [55,56] and model gene regulatory networks [57,58].

Super-essential reactions can also help explain that the incidence of metabolic innovation rises with the number of transferred reactions (Figure 1a, main text). As we showed in the main text, addition of a single reaction is usually sufficient to cause metabolic innovation. By increasing the number of transferred reactions, the probability increases that at least one highly superessential reaction is transferred, and so the incidence of metabolic innovation increases.

In addition, the transfer of superessential reactions can help explain that increasing genotypic distance between donor and recipient decreases the incidence of metabolic innovation (figure 2a, main text). Since the number of highly superessential reactions is limited [59], increasing the genotypic distance between donor and recipient decreases the fraction of such reactions that are not already in the recipient. In consequence, the incidence of innovative offspring decreases as well.

Superessential reactions can also help explain why the incidence of innovation increases with increasing parental phenotypic diversity $\Delta P$ – an increasing number of carbon sources on which one but not the other parent is viable. Any phenotypic difference between parents must be caused by the set of $D$ reactions that are not shared between the parents. As $\Delta P$ increases, an increasing number of these non-shared reactions would be involved in viability on at least one of the carbon sources on which the parents are viable, and such reactions tend to have a higher super-essentiality index [59]. These are also the reactions that will lead to innovation when affected by a recombination event (figure S7). Therefore, parents with higher $\Delta P$ are expected to have a higher fraction of exchangeable reactions with high super essentiality index (Figure 2d in main text), and consequently higher fraction of innovative offspring (Figure 2c in main text).

Finally, with increasing phenotypic complexity ||P|| – the number of carbon sources on which a metabolic network is viable – of parental metabolic networks with the same phenotype the incidence of innovation by recombination decreases. To explain this pattern, consider two genotypically distinct metabolic networks with the same phenotype. Their non-shared reactions are less likely to be essential for viability than their shared reactions, and so the superessentiality index of the non-shared reactions is expected to be low. As ||P|| increases, the fraction of non-shared reactions

with high superessentiality index is expected to decrease further, exactly as we observed (Figures 2f (main text) and S13d), which leads to a lower incidence of innovation (Figures 2e (main text) and S13a).

**Supplementary References:**


1. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. 2001 In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* 19, 125–30. (doi:10.1038/84379)
2. Edwards, J. S. & Palsson, B. O. 1999 Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J. Biol. Chem.* 274, 17410–6.
3. Lewis, N. E., Nagarajan, H. & Palsson, B. O. 2012 Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10, 291–305. (doi:10.1038/nrmicro2737)
4. Feist, A. M. & Palsson, B. Ø. 2008 The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat. Biotechnol.* 26, 659–67. (doi:10.1038/nbt1401)
5. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. 2009 Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* 5, 320. (doi:10.1038/msb.2009.77)
6. McCloskey, D., Palsson, B. Ø. & Feist, A. M. 2013 Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* 9, 661. (doi:10.1038/msb.2013.18)
7. Fondi, M. & Liò, P. 2015 Genome-scale metabolic network reconstruction. *Methods Mol. Biol.* 1231, 233–56. (doi:10.1007/978-1-4939-1720-4_15)
8. Matias Rodrigues, J. F. & Wagner, A. 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* 5, e1000613. (doi:10.1371/journal.pcbi.1000613)
9. Goto, S., Nishioka, T. & Kanehisa, M. 2000 LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* 28, 380–2.
10. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. 2010 KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–60. (doi:10.1093/nar/gkp896)
11. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. & Hirakawa, M. 2006 From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 34, D354–7. (doi:10.1093/nar/gkj102)
12. Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. & Palsson, B. Ø. 2007 A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121. (doi:10.1038/msb4100155)

13. Orth, J. D., Thiele, I. & Palsson, B. Ø. 2010 What is flux balance analysis? *Nat. Biotechnol.* 28, 245–8. (doi:10.1038/nbt.1614)

14. Kauffman, K. J., Prakash, P. & Edwards, J. S. 2003 Advances in flux balance analysis. *Curr. Opin. Biotechnol.* 14, 491–6.

15. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. 2002 LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* 30, 402–4.

16. Kanehisa, M. & Goto, S. 2000 KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

17. Lercher, M. J. & Pál, C. 2008 Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* 25, 559–67. (doi:10.1093/molbev/msm283)

18. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. 2002 Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420, 186–9. (doi:10.1038/nature01149)

19. Vieira-Silva, S. & Rocha, E. P. C. 2010 The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6, e1000808. (doi:10.1371/journal.pgen.1000808)

20. Cox, R. A. 2004 Quantitative relationships for specific growth rates and macromolecular compositions of Mycobacterium tuberculosis, Streptomyces coelicolor A3(2) and Escherichia coli B/r: an integrative theoretical approach. *Microbiology* 150, 1413–26.

21. Kirschner, D. & Marino, S. 2005 Mycobacterium tuberculosis as viewed through a computer. *Trends Microbiol.* 13, 206–11. (doi:10.1016/j.tim.2005.03.005)

22. Fong, S. S. & Palsson, B. Ø. 2004 Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36, 1056–8. (doi:10.1038/ng1432)

23. Fong, S. S., Marciniak, J. Y. & Palsson, B. O. 2003 Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* 185, 6400–6408. (doi:10.1128/JB.185.21.6400-6408.2003)

24. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. 2010 Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* 4, 30. (doi:10.1186/1752-0509-4-30)

25. Burgard, A. P., Nikolaev, E. V, Schilling, C. H. & Maranas, C. D. 2004 Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* 14, 301–12. (doi:10.1101/gr.1926504)

26. Thomas, C. M. & Nielsen, K. M. 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–21. (doi:10.1038/nrmicro1234)

27. Guttman, D. S. & Dykhuizen, D. E. 1994 Clonal divergence in Escherichia coli as a result of recombination, not mutation. *Science* 266, 1380–3.

28. Feil, E. J. et al. 2001 Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. U. S. A.* 98, 182–7. (doi:10.1073/pnas.98.1.182)

29. Whitaker, R. J., Grogan, D. W. & Taylor, J. W. 2005 Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. *Mol. Biol. Evol.* 22, 2354–61. (doi:10.1093/molbev/msi233)

30. Ochman, H., Lawrence, J. G. & Groisman, E. A. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* 405, 299–304. (doi:10.1038/35012500)

31. Pál, C., Papp, B. & Lercher, M. J. 2005 Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37, 1372–5. (doi:10.1038/ng1686)

32. Fraser, C., Hanage, W. P. & Spratt, B. G. 2007 Recombination and the nature of bacterial speciation. *Science* 315, 476–80. (doi:10.1126/science.1127573)

33. Majewski, J., Zawadzki, P., Pickerill, P., Cohan, F. M. & Dowson, C. G. 2000 Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. *J. Bacteriol.* 182, 1016–23.

34. Kowalczykowski, S. C., Dixon, D. A., Eggleston, A. K., Lauder, S. D. & Rehrauer, W. M. 1994 Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.* 58, 401–65.

35. Mira, A., Ochman, H. & Moran, N. A. 2001 Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17, 589–96.

36. Kuo, C.-H. & Ochman, H. 2009 The fate of new bacterial genes. *FEMS Microbiol. Rev.* 33, 38–43. (doi:10.1111/j.1574-6976.2008.00140.x)

37. Bork, P. & Doolittle, R. F. 1992 Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci.* 89, 8990–8994. (doi:10.1073/pnas.89.19.8990)

38. Inagaki, Y., Susko, E. & Roger, A. J. 2006 Recombination between elongation factor 1 genes from distantly related archaeal lineages. *Proc. Natl. Acad. Sci.* 103, 4528–4533. (doi:10.1073/pnas.0600744103)

39. Hartl, D. L., Lozovskaya, E. R. & Lawrence, J. G. 1992 Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica* 86, 47–53.

40. Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K. & Nagashima, K. V 2001 Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J. Mol. Evol.* 52, 333–41. (doi:10.1007/s002390010163)

41. Omelchenko, M. V, Makarova, K. S., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V 2003 Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* 4, R55. (doi:10.1186/gb-2003-4-9-r55)

42. Chan, C. X., Beiko, R. G., Darling, A. E. & Ragan, M. A. 2010 Lateral Transfer of Genes and Gene Fragments in Prokaryotes. *Genome Biol. Evol.* 1, 429–438. (doi:10.1093/gbe/evp044)

43. Denamur, E. et al. 2000 Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell* 103, 711–21.

44. Lin, C. H., Bourque, G. & Tan, P. 2008 A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol. Biol. Evol.* 25, 549–58. (doi:10.1093/molbev/msm282)

45. Didelot, X., Achtman, M., Parkhill, J., Thomson, N. R. & Falush, D. 2007 A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res.* 17, 61–8. (doi:10.1101/gr.5512906)

46. King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O. & Lewis, N. E. 2015 BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* , gkv1049–. (doi:10.1093/nar/gkv1049)

47. Gelius-Dietrich, G., Desouki, A. A., Fritzemeier, C. J. & Lercher, M. J. 2013 Sybil--efficient constraint-based modelling in R. *BMC Syst. Biol.* 7, 125. (doi:10.1186/1752-0509-7-125)

48. Tatusova, T., Ciufo, S., Fedorov, B., O'Neill, K. & Tolstoy, I. 2014 RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* 42, D553–9. (doi:10.1093/nar/gkt1274)

49. Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B. & Stevens, R. L. 2010 High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–82. (doi:10.1038/nbt.1672)

50. Suyama, M., Torrents, D. & Bork, P. 2006 PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–12. (doi:10.1093/nar/gkl315)

51. Zawadzki, P., Roberts, M. S. & Cohan, F. M. 1995 The log-linear relationship between sexual isolation and sequence divergence in Bacillus transformation is robust. *Genetics* 140, 917–32.

52. Jeltsch, A. 2003 Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* 317, 13–6.

53. Zahrt, T. C. & Maloy, S. 1997 Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between Salmonella typhimurium and Salmonella typhi. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9786–91.

54. Barve, A. & Wagner, A. 2013 A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* 500, 203–6. (doi:10.1038/nature12301)

55. Cui, Y., Wong, W. H., Bornberg-Bauer, E. & Chan, H. S. 2002 Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. *Proc. Natl. Acad. Sci. U. S. A.* 99, 809–14. (doi:10.1073/pnas.022240299)

56. Drummond, D. A., Silberg, J. J., Meyer, M. M., Wilke, C. O. & Arnold, F. H. 2005 On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. U. S. A.* 102, 5380–5. (doi:10.1073/pnas.0500729102)

57. Martin, O. C. & Wagner, A. 2009 Effects of recombination on complex regulatory circuits. *Genetics* 183, 673–84, 1SI–8SI. (doi:10.1534/genetics.109.104174)

58. Wagner, A. 2011 The low cost of recombination in creating novel phenotypes: Recombination can create new phenotypes while disrupting well-adapted phenotypes much less than mutation. *Bioessays* 33, 636–46. (doi:10.1002/bies.201100027)

59. Barve, A., Rodrigues, J. F. M. & Wagner, A. 2012 Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 109, E1121–30. (doi:10.1073/pnas.1113065109)

## Supplementary figures:

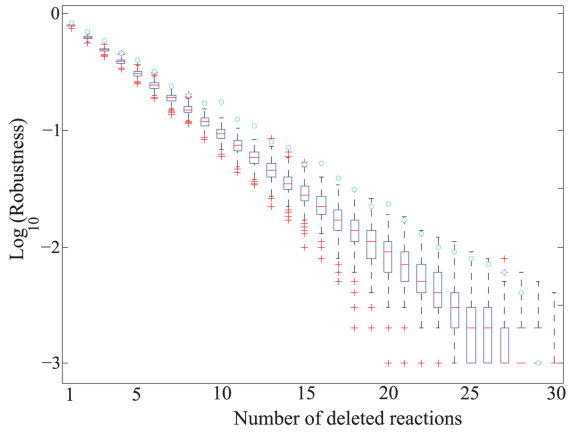



**Figure S1:** Distribution of the fraction of randomly sampled viable metabolic networks (boxes) that retain viability on glucose (*y*-axis, note the logarithmic scale) as compared with that of *E. coli* (cyan circles) after deleting a given number of reactions (*x*-axis). All boxes span the 25-th to 75-th percentile. Horizontal bars in a box indicate the median, and whiskers indicate maxima and minima. Red asterisks indicate outliers.

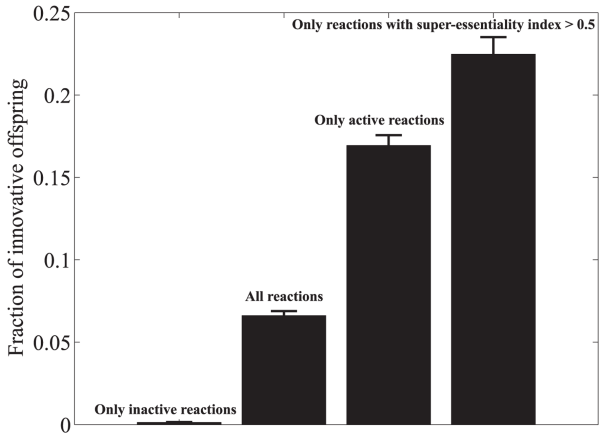



**Figure S2:** Mean (bar), and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$) generated from 1000 random parental metabolic networks viable only on glucose with a fixed genotype distance *D=100*, by adding 5 randomly chosen i) inactive (blocked), ii) active reactions, iii) highly superessential, and iv) mixed (including all types) reactions from a donor metabolic network to the recipient, followed by deleting 5 randomly chosen reactions from the recipient metabolic network.

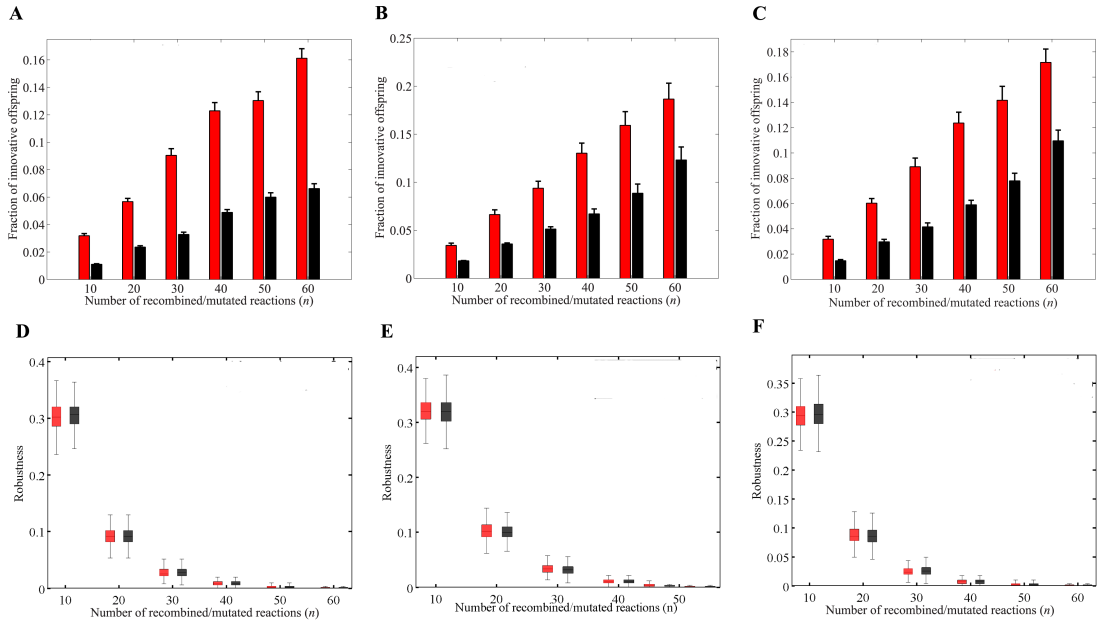



**Figure S3:** Vertical axes in panels **(a), (b),** and **(c)** show mean (bars) and standard error (vertical lines) of the fraction of innovative offspring ($f_{innov}$) generated by recombination (red) versus mutation (black), as a function of the number of reaction changes ($n$, $x$-axis) among those offspring retaining viability respectively on (a) glucose, (b) acetate, and (c) acetate. Parental metabolic network pairs are sampled based on the (a) second, (b) second, and (c) first approach (See texts S1e, and S2). Vertical axes in panels **(d), (e),** and **(f)** show recombinational robustness (red) versus mutational robustness (black) , that are defined as the fraction of recombinant (or mutant) offspring retaining viability respectively on (d) glucose, (e) acetate, and (f) acetate. Parental metabolic network pairs are sampled based on the (d), second (e) second, and (f) first approach (See texts S1f, and S2). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

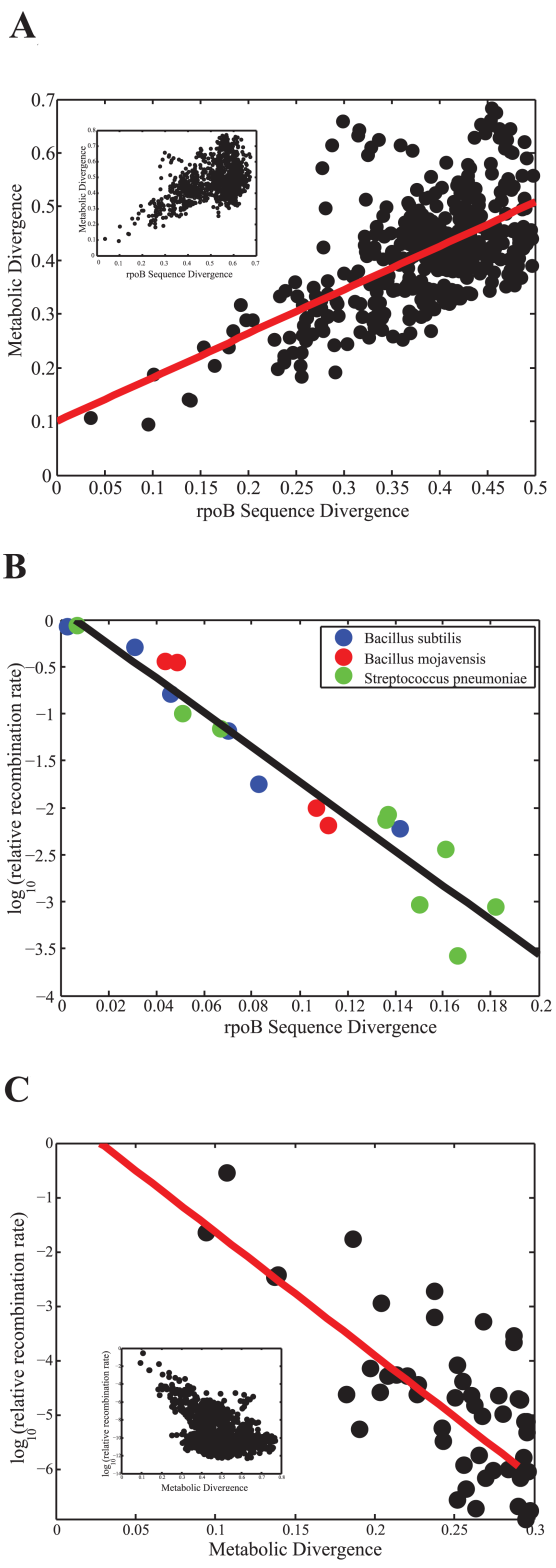



**Figure S4: Relative recombination rates.** **(a)** Metabolic divergence, defined as the normalized Hamming distance between the genotype vectors of two metabolic networks, is correlated (Pearson's $r=0.60$, $P<10^{-40}$) with sequence divergence, defined as the normalized Hamming distance between rpoB (RNA polymerase) sequences of the corresponding pair of species. Each point corresponds to one of $\binom{51}{2}$ possible species pairs chosen from 51 distinct species (inset) whose pairwise rpoB sequence divergence lies below 0.5 [49]. **(b)** Relative rate of recombination, for a range of related donor species as a function of sequence divergence for a variety of bacterial recipients: *Bacillus subtilis* (blue), *Bacillus mojavensis* (red), and *Streptococcus pneumoniae* (green). The best log-linear fit is shown (black line), with an intercept of 0.11 and a slope of -18.40. Data is based on [50,51]. **(c)** Relative recombination rate (logarithmic scale, *y*-axis) as a function of metabolic divergence (*x*-axis) for metabolic network pairs with metabolic divergence lower than 0.3, chosen among the set of all possible metabolic network pairs (inset). The red line is the result of a linear regression with a regression coefficient of -22.57, and an intercept of 0.62. Data in (c) is based on the linear relationship from (b), and the metabolic divergence of (a).

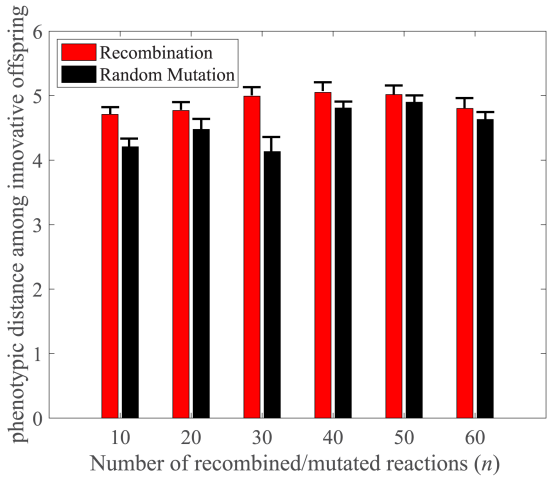



**Figure S5: Phenotypic diversity among innovative offspring.** Vertical axis shows mean (bars) and standard error (vertical lines) of the phenotypic distance among all pairs of innovative offspring generated by recombination (red) versus mutation (black), as a function of the number of reaction changes ($n$, $x$-axis). Phenotypic distance ($\Delta P$) between a given pair of innovative offspring is measured as the number of carbon sources on which only one offspring but not the other is viable on.

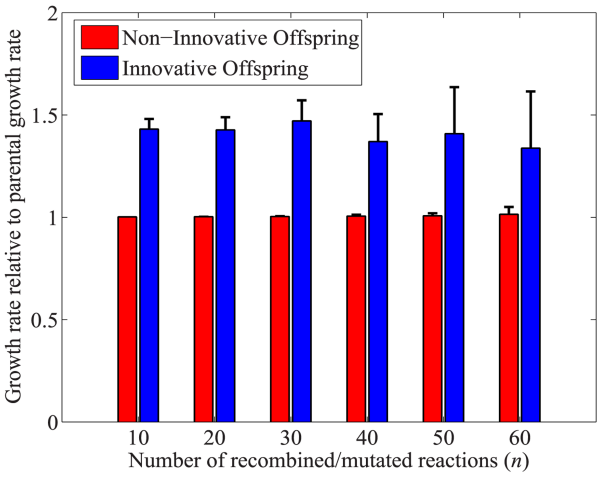



**Figure S6:** Mean (bar) and standard error (vertical line) of the biomass growth flux of innovative offspring (blue), and non-innovative offspring (red), divided by the parental growth rate, as a function of the number of recombined reactions ($n$). For this analysis, we created 1000 random metabolic network pairs viable only on glucose and with a difference in growth rate less than 0.25 percent. Regardless of $n$, the relative growth rate of non-innovative offspring is approximately equal to one, meaning that their growth rate is equal to the parental growth rate. In contrast, the relative growth rate for innovative offspring exceeds 1.4 for all $n$, and so innovative offspring even on the original carbon source can grow more than 40 percent faster than their parents.

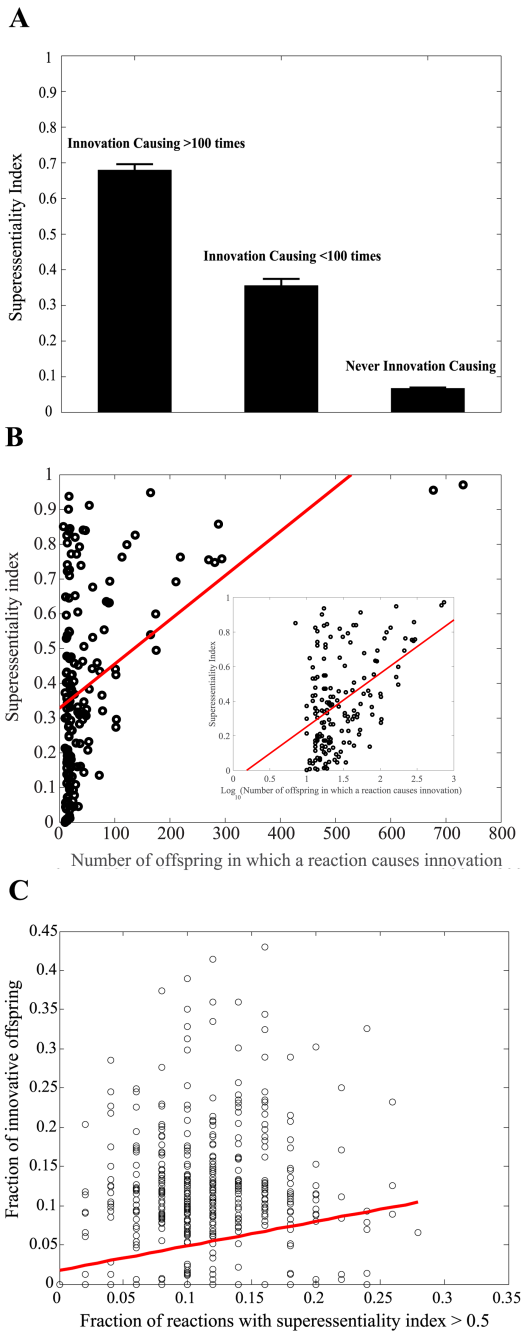



**Figure S7: Superessential reactions and metabolic innovation. (a)** Mean (bars) and standard error (vertical lines) of the superessentiality index of reactions that (i) cause innovation in more than 100 innovative offspring (left), (ii) cause innovation in fewer than 100 innovative offspring (middle), and (iii) never cause innovation (right). **(b)** Scatterplot of superessentiality index ($I_{SE}$; y-axis) versus number of innovations (x-axis) caused by innovation-causing reactions (positive correlation: (Pearson's $r=0.47$, $P<10^{-9}$)). Horizontal axis in the inset is shown in logarithmic scale to improve visual clarity. **(c)** The fraction of innovative recombinant offspring ($f_{innov}$, $y$-axis) is significantly correlated (Pearson's $r=0.18$, $P<10^{-8}$) with the fraction of reactions with superessentiality index higher than 0.5 ($f_{super}$, $x$-axis) among reactions that can potentially be transferred from donor to the recipient.

When studying metabolic innovation, it is important to distinguish two classes of reactions with high superessentiality index. The first comprises reactions with superessentiality index $I_{SE}$=1, which are needed in all viable metabolic networks [59]. These reactions are crucial for retaining viability on parental carbon sources, but they play no role in metabolic innovation, because all metabolic networks must have them. The second class includes reactions where $0.5< I_{SE} <1$. These reactions are less crucial for retaining viability on parental carbon sources, but important for gaining viability on novel carbon sources. They can be absent in some metabolic networks, because metabolic pathways that by-pass them exist, which means that they can be involved in recombinational exchange, and thus in the origin of novel phenotypes. Our analysis above highlights the special importance of these reactions for metabolic innovation.

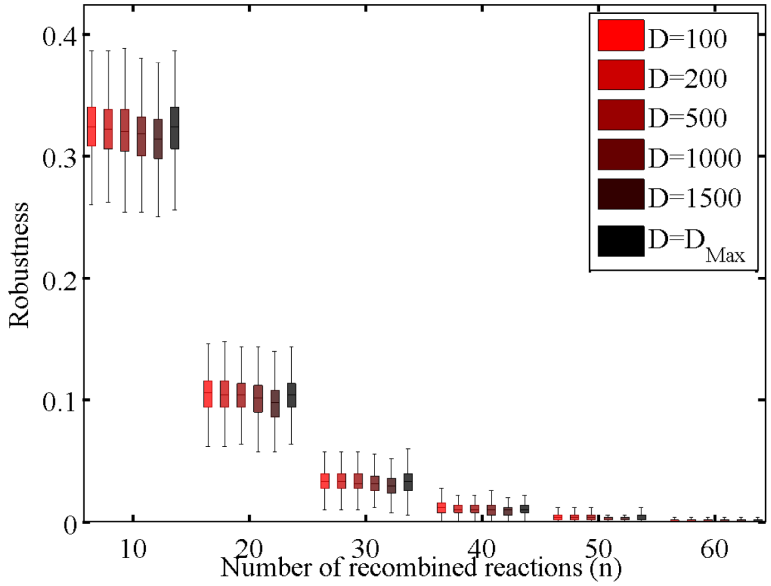



**Figure S8: Effect of parental genotypic diversity on recombinational robustness.** The vertical axis shows recombinational robustness, that is, the fraction of offspring that retain viability on glucose and that are generated by recombination between parental metabolic networks with genotypic distance (*D*), where *D* is color-coded according to the legend. The horizontal axis shows the number of recombined reactions (*n*). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

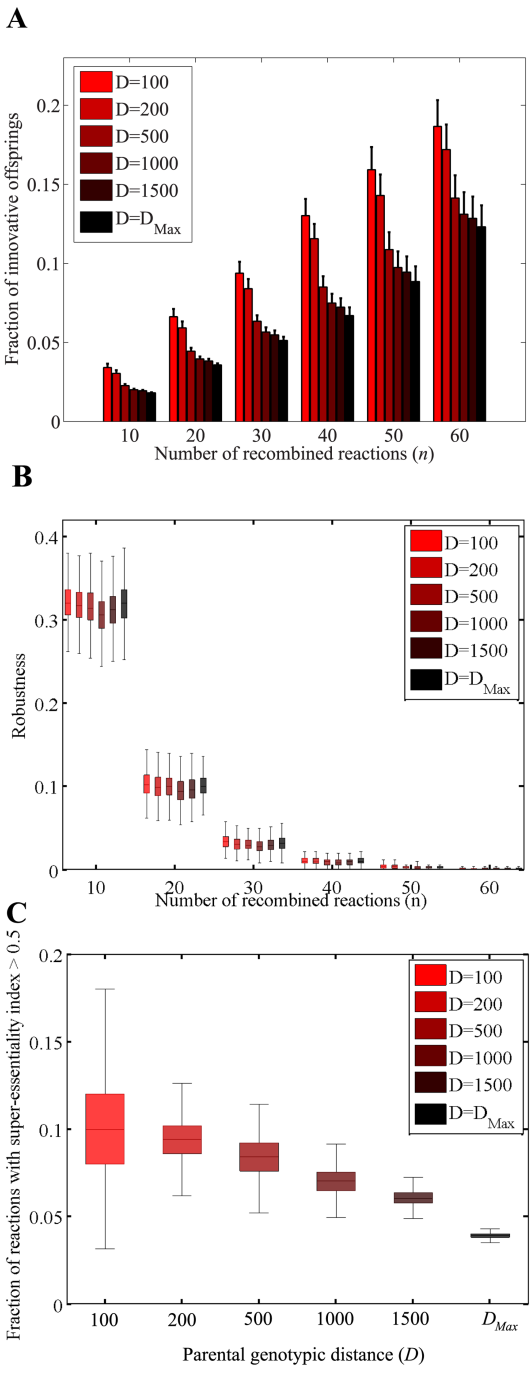



**Figure S9: Effect of parental genotypic diversity on recombinational innovation and robustness (parental metabolic networks are required to be viable on acetate). (a)** the mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$), generated by recombination between parental metabolic networks viable on acetate with genotypic distance ($D$), where $D$ is color-coded according to the legend. The horizontal axis shows the number of recombined reactions ($n$). **(b)** The vertical axis shows recombinational robustness, that is, the fraction of offspring that retain viability on acetate and are generated by recombination between parental metabolic networks with genotypic distance ($D$), where $D$ is color-coded according to the legend of panel (a). The horizontal axis shows the number of recombined reactions ($n$). **(c)** The fraction of reactions with superessentiality index higher than 0.5 ($f_{super}$, $x$-axis) among reactions that can potentially be transferred from the parental donor to the recipient metabolic network, with genotypic distance ($D$, $x$-axis). Note that parental metabolic networks are required to be viable on acetate instead of glucose. All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

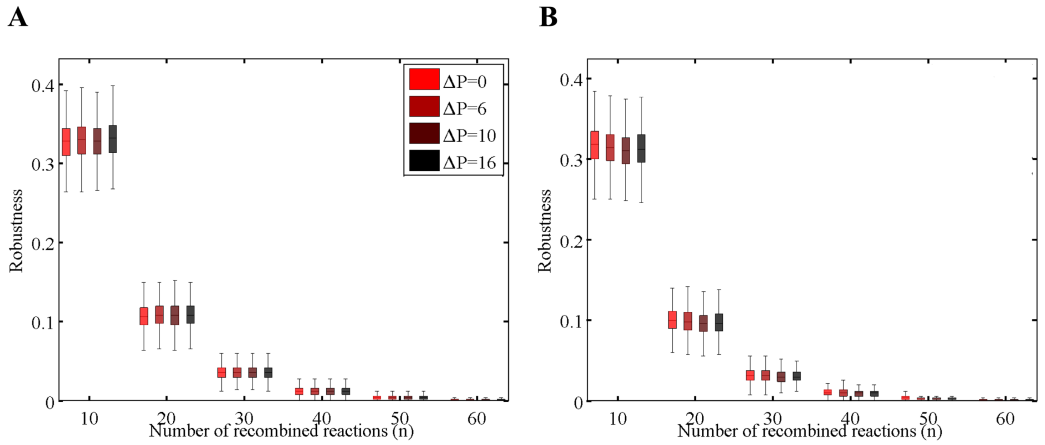



**Figure S10: Effect of parental phenotypic diversity on recombinational robustness.** The vertical axes show **(a)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in **(b)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity *ΔP* (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions (*n*). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

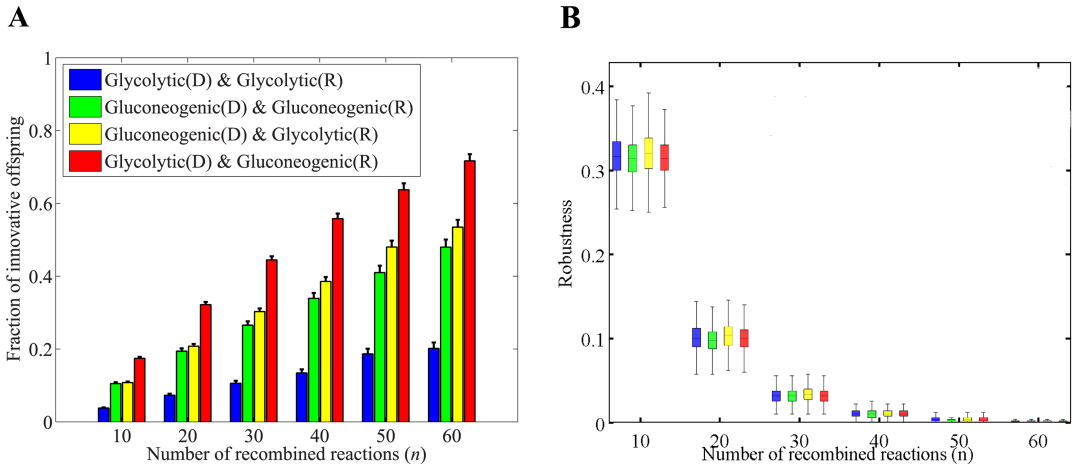



**Figure S11: Effect of parental carbon source classes on metabolic innovation.** The vertical axes in **(a)** show the mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$), and in **(b)** the fraction of recombinant offspring retaining viability on glucose (for those recipients viable on glycolytic carbon sources) or acetate (for those recipients viable on gluconeogenic carbon sources). The horizontal axes show the number of recombined reactions (*n*). For this analysis we generated offspring by recombination between parental metabolic networks in which (i) the donor was viable on 5 glycolytic carbon sources and the recipient was viable on 5 other glycolytic carbon sources (blue), (ii) the donor was viable on 5 gluconeogenic carbon sources and the recipient was

viable on 5 other gluconeogenic carbon sources (green), (iii) the donor was viable on 5 gluconeogenic carbon sources and the recipient was viable on 5 glycolytic carbon sources (yellow), and (iv) the donor was viable on 5 glycolytic carbon sources and the recipient was viable on 5 gluconeogenic carbon sources (red). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

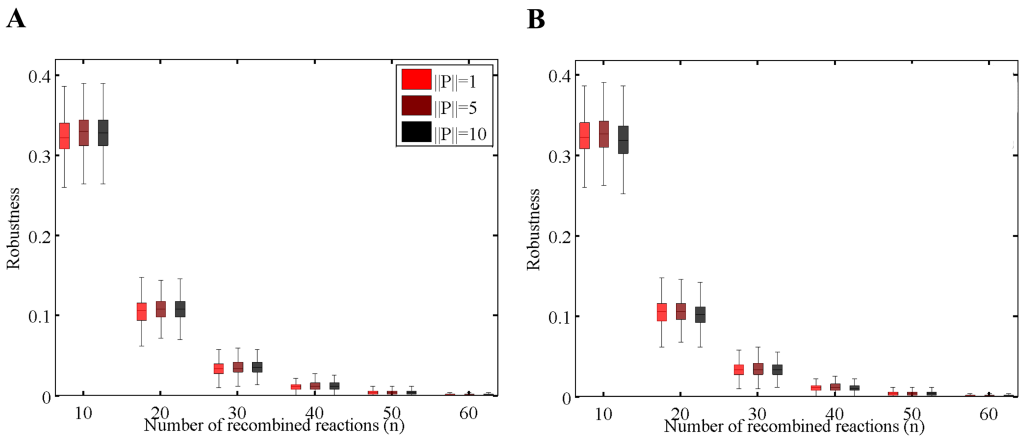



**Figure S12: Effect of parental phenotypic complexity on metabolic robustness (for $||P|| \leq 10$, and $\Delta P=0$)**. The vertical axes show (**a)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in (**b)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity $||P||$ (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions ($n$). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima

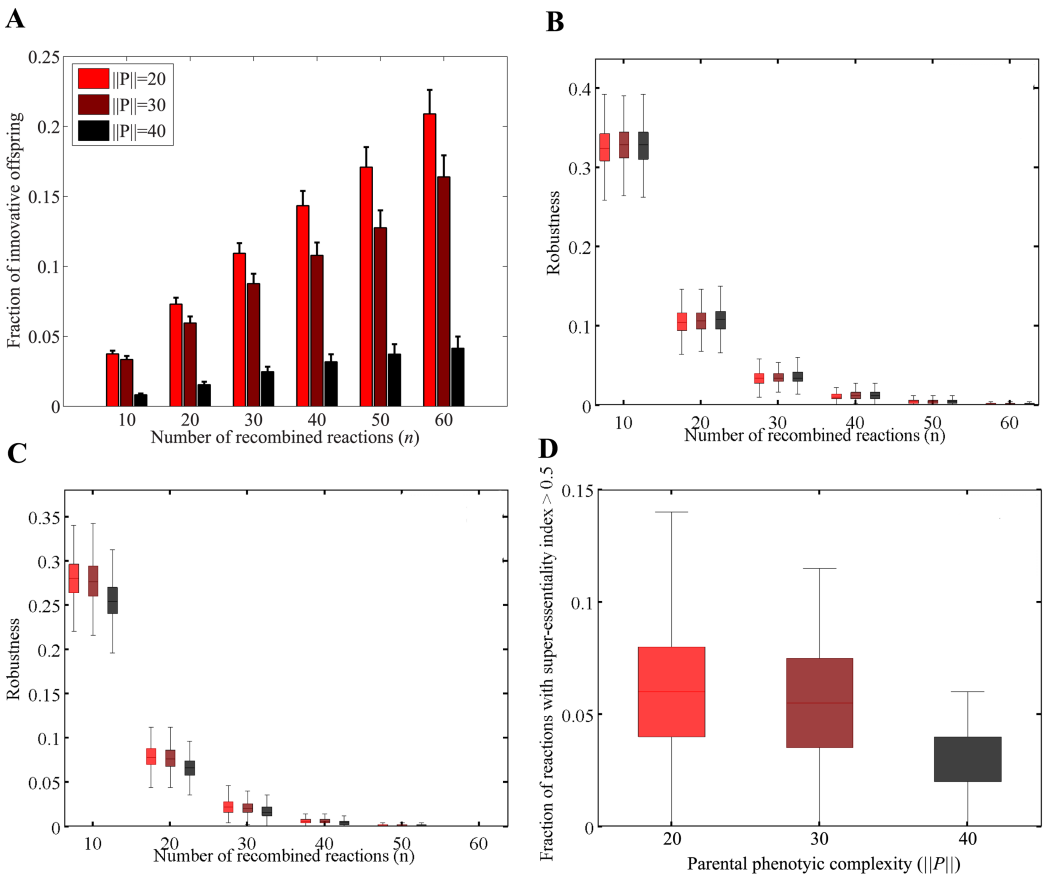



**Figure S13: Effect of parental phenotypic complexity on metabolic innovation (for $||P|| > 10$, and *ΔP=10*)**. The vertical axes show in (**a)** the mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$), in (**b)** the fraction of recombinant offspring retaining viability on glucose (i.e. robustness), and in (**c)** the fraction of recombinant offspring retaining viability on all carbon sources (not only glucose) on which the corresponding recipient parental metabolic network is viable. Offspring were generated by recombination between parental metabolic networks with phenotypic complexity ||*P*|| (color-coded as shown in the legend in panel (a)). The horizontal axes show the number of recombined reactions (*n*). (**d)** Distribution of the fraction of reactions with superessentiality index exceeding 0.5 (*y*-axis) among the reactions that can potentially be transferred from the parental donor metabolic network to the recipient, with phenotypic complexity (||*P*||, *x*-ais). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

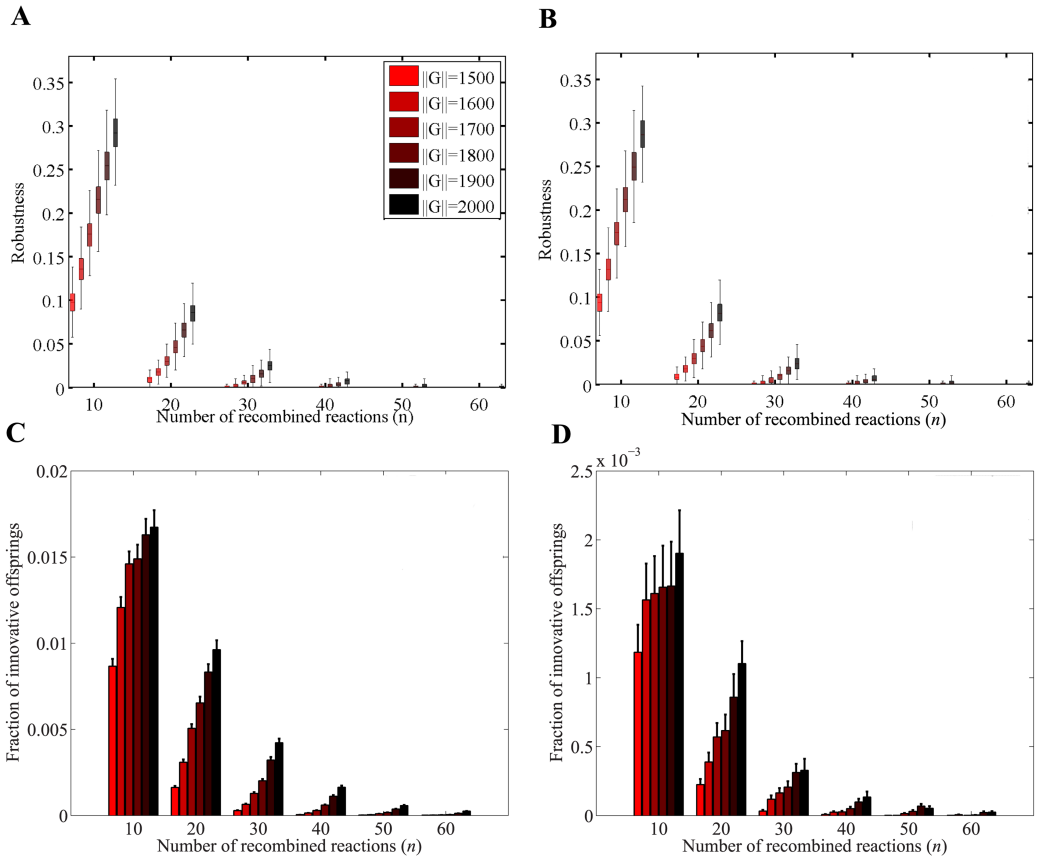



**Figure S14: Effect of genotypic complexity (metabolic network size *(||G||)*) on recombinational innovation.** Vertical axes in panels **(a)**, and **(b)** show the fraction of recombinant offspring retaining viability (i.e. robustness, vertical axis), on (a) glucose, and (b) acetate, are shown as a function of the number of recombined reactions (*n*). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima. Panels **(c)**, and **(d)** show mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$), generated by recombination between parental metabolic networks required to be viable on (c) glucose, and (d) acetate, with size *(||G||)* color-coded as in the legend, are shown as a function of the number of recombined reactions (*n*).

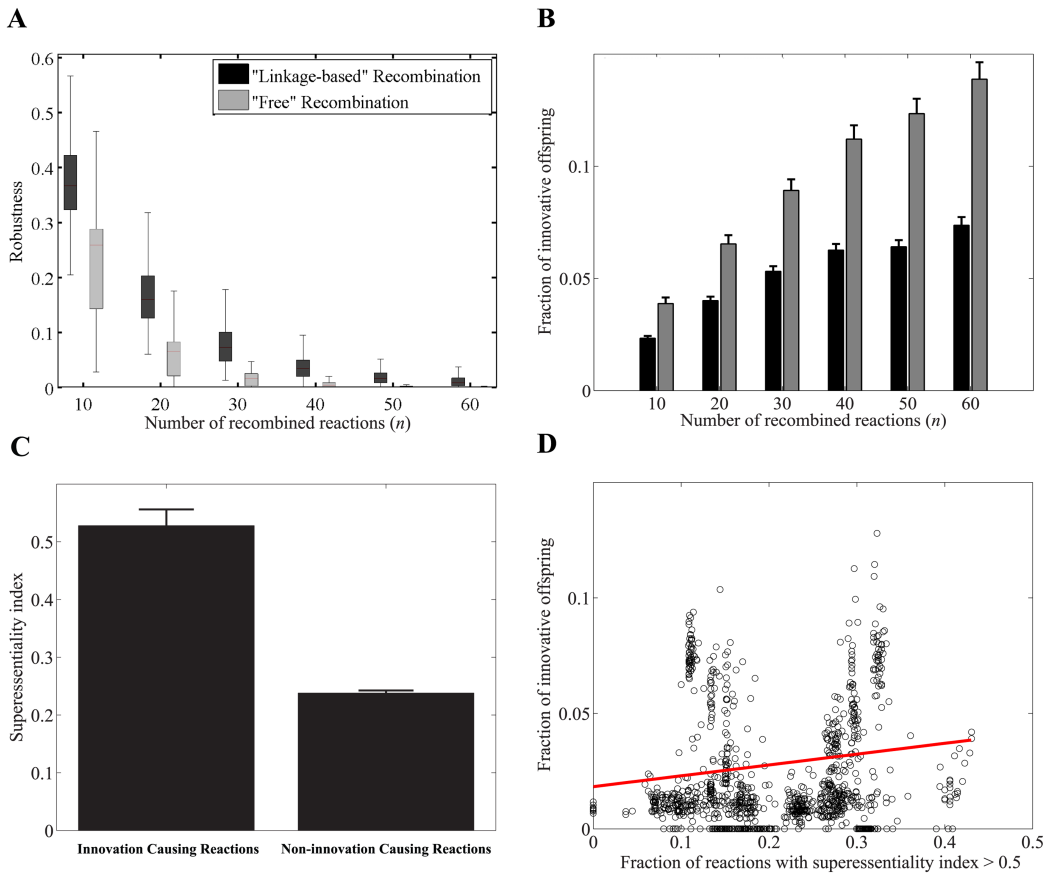



**Figure S15: (a)** Fraction of robust recombinant offspring, i.e., offspring retaining viability on all the carbon sources that the recipient parental metabolic network is viable on (*y*-axis), as a function of the number of recombined reactions (x-axis). Offspring were generated by i) linkage-based recombination between prokaryotic metabolic networks (black), and ii) free recombination between prokaryotic metabolic networks (gray). All boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima**. (b)** Mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$) generated by (i) linkage-based recombination (black), and (ii) free recombination between prokaryotic metabolic networks (gray)**. (c)** Mean (bars) and standard error (vertical lines) of the superessentiality index of reactions that cause innovation (left) as compared with those never causing innovation (right). **(d)** The fraction of innovative recombinant offspring ($f_{innov}$, *y*-axis) is significantly associated (Pearson's $r=0.13$, $P<10^{-5}$) with the fraction of reactions with superessentiality index higher than 0.5 (*x*-axis) among the set of reactions that can potentially be transferred from the parental donor to the recipient metabolic network.

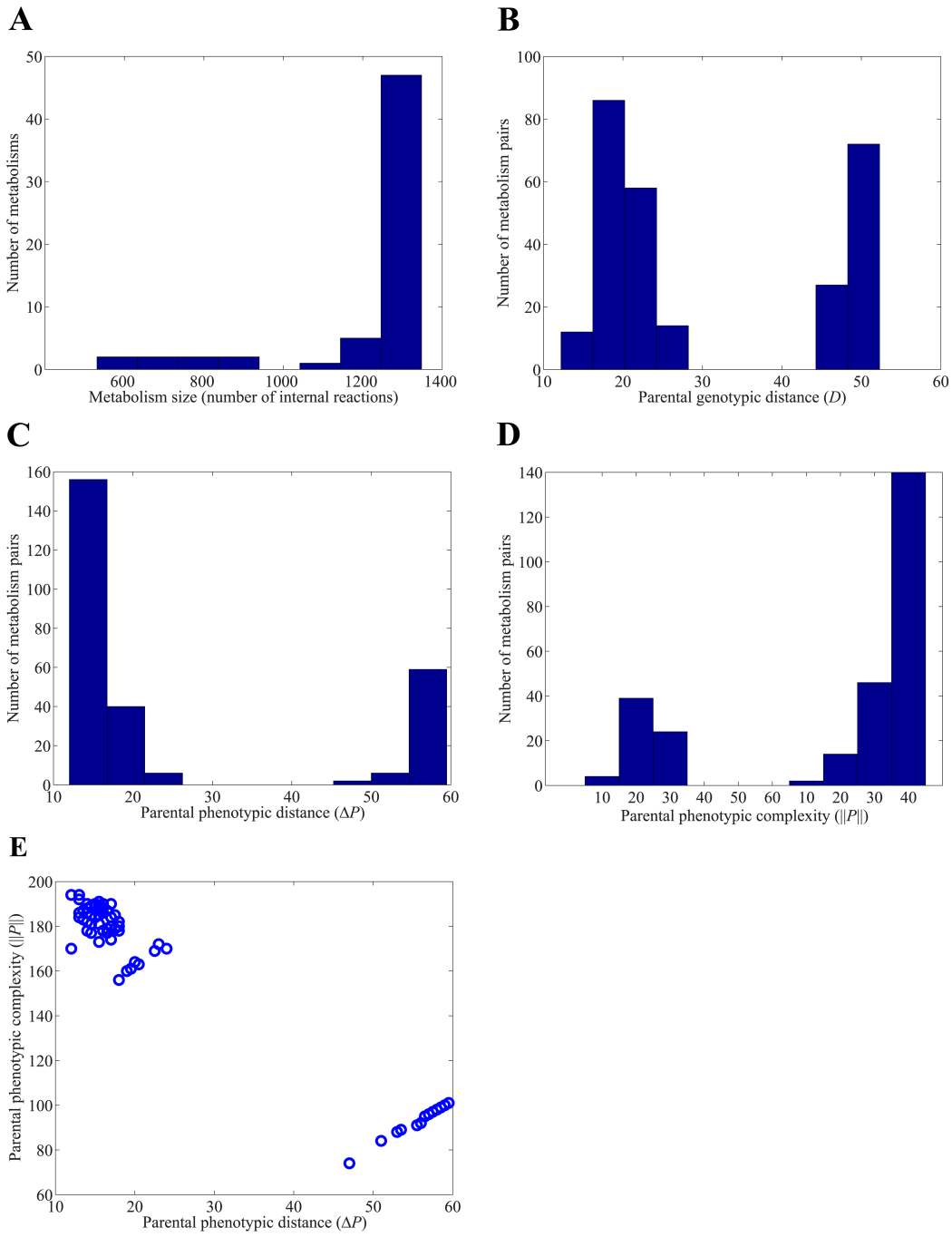



**Figure S16: Distribution of parental genotypic and phenotypic features among prokaryotic metabolic networks. (a)** Histogram of the number of metabolic networks with a given metabolic network size (approximated by the number of internal reactions) specified on the *x*-axis. Vertical axes in panels **(b)**, **(c)**, and **(d)** show the number of parental metabolic network pairs with a given b) genotypic distance (*D*), c) phenotypic distance (*ΔP*), and d) phenotypic complexity (||*P*||), as specified on the *x*-axes. **(e)** Each circle represents a given parental metabolic network pairs with a given phenotypic distance (*ΔP*, *x*-axis), and phenotypic complexity (||*P*||, *y*-axis).

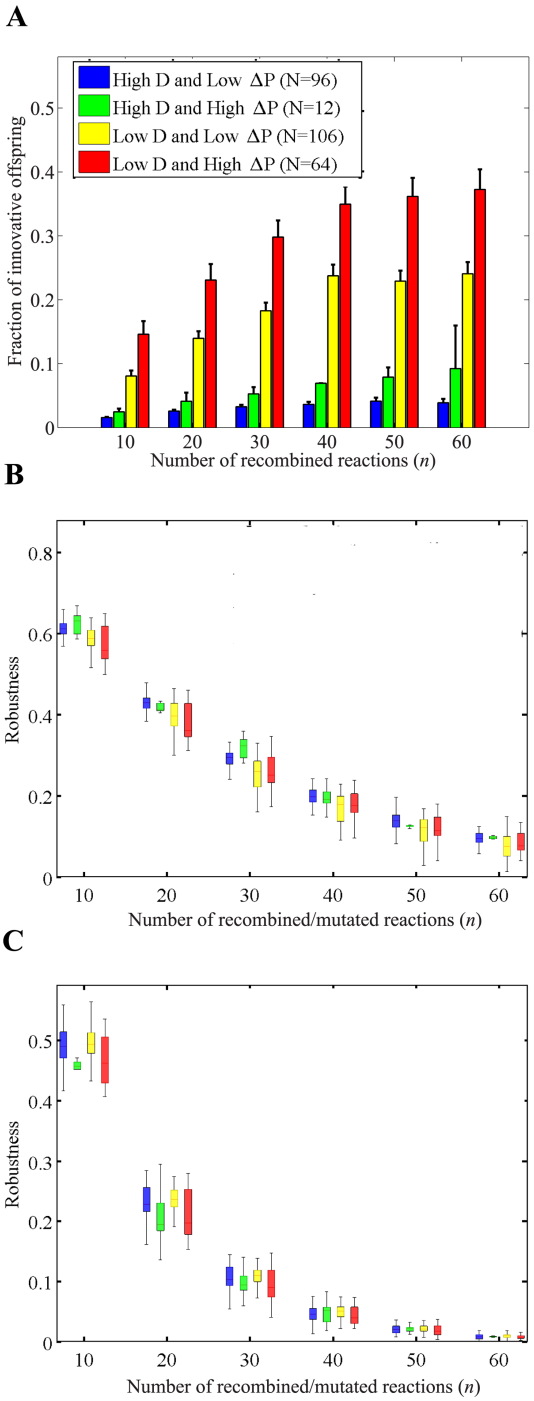



**Figure S17:** The vertical axes show **(a)** mean (bar) and standard error (vertical line) of the fraction of innovative offspring ($f_{innov}$), **(b)** robustness to linkage-based, and **(c)** robustness to "free"recombination. Here we define robustness as the fraction of recombinant offspring retaining viability on at least one of the carbon source(s) on which the parental recipient metabolic networks are viable. Offspring are generated by recombination between prokaryotic parental metabolic networks with i) high genotypic distance ($D>40$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (blue, $N=96$ parental pairs), ii) high genotypic distance ($D>40$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (green, $N=12$ parental pairs), iii) low genotypic distance ($D<30$), low phenotypic distance ($\Delta P<30$), and high phenotypic complexity ($||P||>60$) (yellow, $N=106$ parental pairs), and iv) low genotypic distance ($D<30$), high phenotypic distance ($\Delta P>40$), and low phenotypic complexity ($||P||<40$) (red, $N=64$ parental pairs).

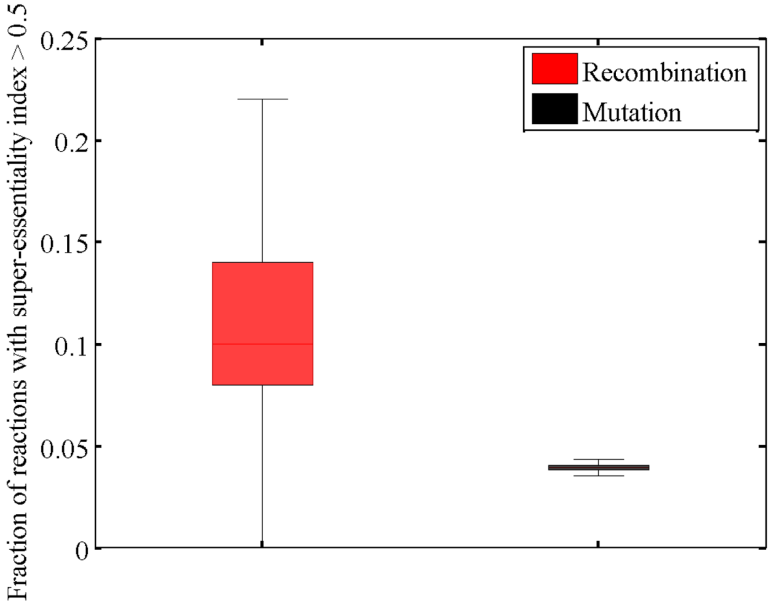



**Figure S18:** The fraction of reactions with superessentiality index higher than 0.5 ($f_{super}$, *x*-axis) among the set of reactions that can potentially be transferred to the recipient metabolic network via recombination (red box) versus random mutation (black box). Boxes span the 25-th to 75-th percentile, horizontal bars in a box indicate the median, and whiskers indicate maxima and minima.

# Chapter 3:

# Genomic organization underlying deletional robustness in bacterial metabolic systems

Sayed-Rzgar Hosseini and Andreas Wagner

The manuscript corresponding to this chapter is under review in PNAS.

## 3.1. Abstract

Large scale DNA deletions and gene loss are pervasive in bacterial genomes. This observation raises the possibility that evolutionary adaptation has altered bacterial genome organization to increase its robustness to large-scale tandem gene deletions. To find out, we systematically analyzed 55 bacterial genome-scale metabolisms, and showed that metabolic gene ordering renders an organism's viability in multiple nutrient environments significantly more robust against tandem multi-gene deletions than expected by chance. This excess robustness is caused by multiple factors, which include the clustering of essential metabolic genes, a greater than expected distance of synthetic lethal non-essential metabolic gene pairs, and the clustering of non-essential metabolic genes. By computationally creating minimal genomes, we show that a non-adaptive origin of such clustering could in principle arise as a passive by-product of bacterial genome growth. However, because genome randomization forces such as translocation and inversion would eventualy erode such clustering, adaptive processes are necessary to sustain it. Our analyses of essential metabolic genes in operons are consistent with the notion that gene deletions are important for this adaptation. Also, we provide evidence that horizontal gene transfer contributes to sustain essential gene clustering. Our observations suggest that the genome organization of bacteria is driven by adaptive processes that provide phenotypic robustness in response to large-scale gene deletions. This robustness may be especially important for bacterial populations that take advantage of gene loss to adapt to new environments.

## 3.2. Significance

From the organismal and the anatomical levels down to the molecular level, all complex biological systems manifest astonishing organization and order that are counter-intuitive and challenging to explain by evolutionary mechanisms. In this study, we focus specifically on one aspect of this biological organization, the arrangement of metabolic genes in bacterial genomes. We show that this organization ensures a substantially higher robustness to large-scale gene deletions than expected from random genomic ordering. We systematically investigate the possible evolutionary mechanisms behind the emergence of such robust organizations. Our analysis provides several lines of evidence indicating that bacteria may have gained a robust genome organization through pervasive gene loss events.

## 3.3. Introduction

Bacterial genomes evolve highly dynamically. On the one hand, they expand through gene gain mechanisms such as horizontal gene transfer (HGT) (1). On the other hand, they contract via large scale gene loss (2). Large-scale gene deletion events were first documented in obligate pathogens and symbionts (3, 4), but later comparative genomic studies showed that they are surprisingly pervasive in bacterial genomes in general (5, 6). Importantly, bacterial genomes experience a well-known general bias towards DNA deletion, that is, genome size reduction events prevail over genome size expansion events (7, 8). Moreover, according to experimental evolution studies, extensive gene loss by large-scale deletions can readily occur on short evolutionary time-scales (9–11).

Does the high incidence of large-scale gene deletions leave evolutionary signatures in bacterial genomes? We hypothesized that bacterial genomes have evolved an organization that provides robustness against the deleterious phenotypic effects of large-scale gene losses. Because large-scale deletional events typically delete multiple contiguous (linked) genes, such a robust genome organization should ensure that a *tandem* deletion of multiple linked genes is on average more tolerable than a deletion of the same number of genes *randomly* drawn from the genome without regard to linkage. In other words, bacterial genomes should be more robust to tandem deletion than random deletion of the same number of genes.

To validate this hypothesis, we focused on metabolic genomes, which encode the enzymes catalyzing the chemical reactions of metabolism. Compared to other biological systems, metabolism is particularly appropriate for such validation, because well-established and experimentally validated computational methods to predict complex phenotypes – especially a cell's viability in specific environments – from genomic information are available (12). What is more, well-annotated genome-scale metabolic networks with information about metabolic genes, reactions, gene-reaction association rules, and the relative genomic order of metabolic genes are available for multiple bacterial genomes (13, 14). Our analysis is based on such information from 55 well-studied bacterial species or strains (14).

## 3.4. Results and discussion

To quantify how metabolic gene order affects the phenotypic robustness of a metabolism to gene deletions, we subjected the metabolic genome of *Escherichia coli* K-12 G1655 to two different kinds of multi-gene deletions. First, in tandem deletions, we deleted a given number of *n* metabolic genes in the order in which they occur in the *E. coli* genome. Second, in random deletions, we deleted *n* randomly chosen metabolic genes irrespective of their order in the genome. More specifically, for every value of *n* between 1 and 50, tandem deletion involved deleting all possible consecutive *n*-tuples of these genes (see methods). For random deletions, we deleted an equivalent number of randomly chosen *n*-tuples of genes. Next, we mapped the eliminated genes in these deletion variants to eliminated reactions in the *E. coli* metabolism and determined the viability of each deletion variant on up to 103 carbon sources using flux balance analysis (12) (table S1). We computed the robustness *R* of the *E. coli* metabolic genome to such gene deletions as the fraction of deletion variants that retain viability on at least one of the carbon sources, either for tandem deletion ($R_{tandem}$) or random deletion ($R_{random}$).

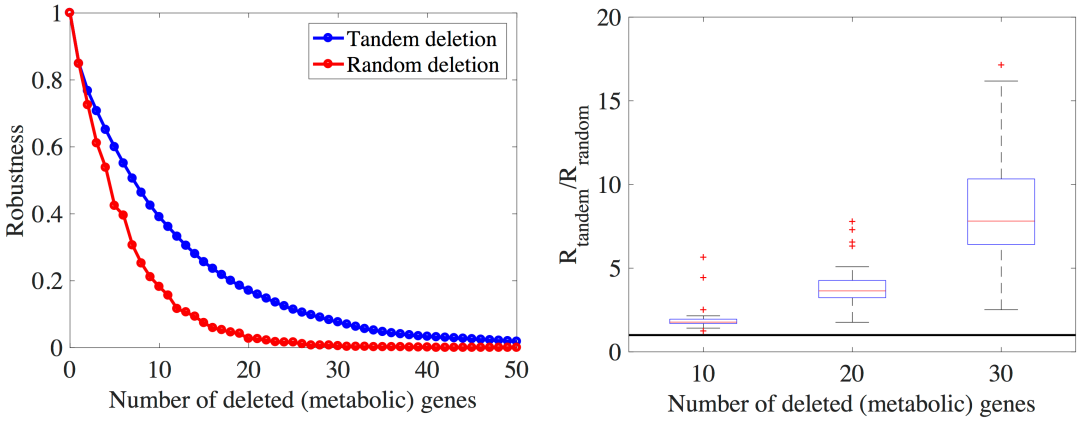



**Figure 1: Robustness to tandem deletion versus random deletion. A)** The vertical axis shows the robustness of *Escherichia coli* K-12 G1655 *(iJO1366)* to tandem (blue) and random (red) deletion of metabolic genes, averaged over all deletional variants we examined, as a function of the number of deleted genes (horizontal axis). **B)** Excess robustness to tandem deletion defined as the ratio of robustness to tandem deletion and robustness to random deletion ($R_{tandem}/R_{random}$), for all 55 bacterial genomes, as a function of the number of deleted genes (horizontal axis). In panels A and B, robustness is defined as the fraction of deletional variants that retain viability on at least one carbon source.

We observed that robustness to tandem deletions is higher for all numbers $n>1$ of deleted genes, and sometimes considerably so (Figure 1a). The same observation holds when we used a more strict definition of robustness; namely the fraction of deletional variants that retain viability on all carbon sources on which the wild type *E. coli* is viable (figure S1). We also repeated this analysis for the 54 other prokaryotic genomes, and observed the same patterns in all of them (See figure S2 for two examples). To quantify by how much robustness to tandem deletions is higher than to random deletions, we computed the ratio $R_{tandem}/R_{random}$, which we call the excess robustness under tandem deletion. For example, for deletions of 20 genes, robustness to tandem deletions is on average 3.63-fold higher than to random deletion (Figure 1b). This excess robustness increases with the number of deleted genes (Figures 1b and S3). In other words, gene order increases in its importance for deletional robustness as deletions become larger. Moreover, by considering robustness based on viability on single individual carbon sources, we observed that robustness to tandem deletion is more conserved among bacterial species or strains than robustness to random deletion (figure S4). We also noted that robustness varied to a greater extent among carbon sources for tandem deletion, than for random deletion (Figures S5 and S6).

Next, we aimed to identify the underlying causes of the excess robustness to tandem deletion. We first wanted to find out whether our observations might be trivially explained by the selfish operon hypothesis (15), which was proposed to explain the clustering of functionally related genes into operons. This hypothesis asserts that the organization of genes into operons is not necessarily beneficial for a host genome, but for the constituent genes, because an operon enables the spreading of its genes to new cells and species by horizontal gene transfer. The hypothesis, which has been criticized before (16, 17), also predicts that horizontally transferred operons would harbor genes with peripheral (i.e. non-essential) metabolic functions (15), such that their deletion would be more tolerable than the deletion of other genes in the genome. Using the DOOR database (18, 19), a comprehensive database for operon information, we showed that the excess tandem robustness we observe is not simply explained by the dispensability of operonic genes as implied by the selfish operon hypothesis (see text S1, figure S7, and table S2).

We then focused on an important class of metabolic genes called *essential genes*. Deletion of an essential metabolic gene alone would be enough for a metabolism to

lose viability in a given environment. These genes and their organization might therefore be important to explain a genome's excess robustness to tandem gene deletions. It has been shown previously that essential genes play a key role in shaping chromosome organization (20), and that they are not uniformly distributed but clustered in bacterial genomes (21, 22). Such clustering can increase the robustness of a genome to tandem multi-gene deletions. If essential genes were distributed uniformly in the genome, each region of genome would have an approximately equal chance to include at least one essential gene, whose deletion would be lethal. In contrast, if essential genes are densely packed in some genomic regions (i.e. clustered), other regions must be depleted of essential genes. Deletions in the latter regions would be non-lethal, such that this genome organization effectively increases robustness to multi-gene deletions (figure S8). Because we have used multiple environments in our analysis, we distinguished two types of essential metabolic genes: *i)* strictly essential genes, which are essential on all carbon sources, and *ii)* conditionally essential genes, which are essential on at least one carbon source. Using Kuiper's test (23) we showed that in the vast majority of the bacterial genomes both types of essential metabolic genes are significantly clustered (see text S2 and tables S3, S4 and S5).

We then asked how these clusters were originally formed in bacterial genomes. The most fundamental question is whether they originated non-adaptively or adaptively, e.g., in response to ongoing large-scale gene deletions? To answer this question, we first examined a simple non-adaptive scenario that ignores the effects of gene deletions. This scenario is inspired by the concept of a minimal genome, which has been used by multiple researchers as a model for the genome of early DNA-based life forms (24–27). In a minimal metabolic genome, no one gene can be removed without destroying viability, so every gene is essential. A minimal genome is basically a single cluster of essential genes. If present-day genomes evolved from minimal genomes largely by the insertion of genes, then the observed present day clustering of essential genes might be a mere remnant of their clustering in the minimal genome, and thus a non-adaptive byproduct of evolutionary genome growth. To validate this hypothesis, we used a previously established algorithm (28) (see methods) that serially deletes individual genes to generate minimal (metabolic) genomes that can sustain life on a given carbon source. We then re-inserted the missing metabolic genes step-by-step in random locations until we had reinstated a genome with the same number and identity of genes

but a different gene order. Applying this method to the genomes of three bacterial species showed that the extent of essential gene clustering is similar to that of the corresponding wild-type genomes (see text S3, and figures S9-S12). Thus, a simple non-adaptive process can in principle explain the extent of essential gene clustering observed in modern genomes.

However, this simplistic model of genome evolution has several limitations. First, although the minimal genome approach is popular (24–27), the minimal genomes it creates may not approximate the genome organization of early cells. Second, genome evolution involves many more processes, including ongoing gene deletions and duplications. A similar analysis that includes such processes shows that gene deletions enhance the clustering of essential genes further, whereas gene duplications reduce it somewhat (see text S3 and figures S9-S12). This implies that adaptive processes in which selective pressure is imposed by large-scale gene deletions can also contribute to the clustering of the essential genes. Thus, the origin of their clustering may not be purely non-adaptive. Finally and most importantly, frequently occuring genome rearrangement processes (29, 30) like translocation and inversion may cause clusters of essential genes to erode (see text S4 and figures S13 and S14). Thus, even if non-adaptive mechanisms may have created essential gene clustering, other mechanisms may be needed to maintain such clustering.

We aimed to detect potential signatures of such adaptive processes. In doing so, we focused first on operons, because operons are continually built, destroyed, and reshuffled in the course of evolution (31). If large-scale gene deletions impose substantial selective pressure on genome organization, then operons that contribute to robustness to gene deletions, because their essential genes are clustered, would be preferentially preserved under deletion pressure. If so, then operons that exist in present-day bacterial genomes should show significantly greater clustering of essential genes than expected by chance.

To find out, we used operonic information for 52 bacterial genomes from the DOOR databse. We first observed that in each of these bacterial genomes approximately 40% and 20% of operons contain at least one conditionally essential metabolic gene and at least one strictly essential metabolic gene, respectively (table S9). Importantly, we observed that essential metabolic genes are more likely to be part of an operon than other metabolic genes (figures 2a and S15, and tables S10 and S11). We then

partitioned each genome's set of strictly essential metabolic genes into operonic genes and non-operonic ones. Using Kuiper's test we quantified the clustering of genes in each group separately. Operonic essential genes are significantly clustered in all bacterial genomes, but non-operonic genes are significantly clustered only in 12 genomes (23.07%; Figure 2b and table S12). The same association exists when we consider genes that are essential for viability on a single carbon source such as glucose (figure S16).

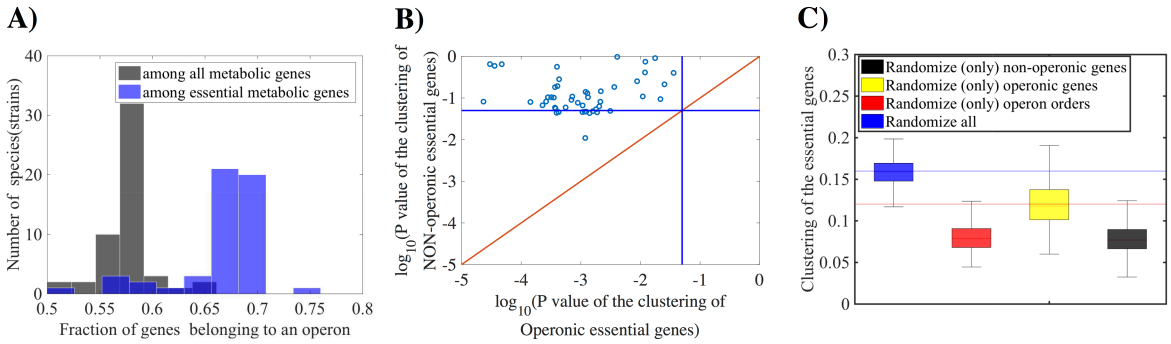



**Figure 2: Operons and essential genes. A)** Histogram of the fraction of all metabolic genes (black), and the fraction of strictly essential metabolic genes (blue) which belong to an operon, based on the 52 species or strains used in this analysis. We consider a metabolic gene as strictly essential, if its deletion results in losing viability on all the carbon sources on which the wild type metabolism is viable. **B)** In this analysis, we subdivided all strictly essential genes in each of 52 metabolic genomes into two groups i) those belonging to an operon and ii) those not belonging to an operon. For each genomes, we determined the extent of gene clustering using the *P*-value generated by Kuiper's test. Each circle in this figure corresponds to a given species or strain. The horizontal and vertical axes show the extent of clustering for genes that are part of an operon and not part of an operon, respectively. The blue lines correspond to a significance threshold of *P*=0.05 ($-log_{10}0.05$), and the red line is the identity line. Note that operonic genes are significantly clustered in more genomes. Whereas none of our 52 genomes show evidence for clustering of essential genes outside operons (at *P*=0.05), essential genes in operons are clustered in 40 of the 52 (76.93%) genomes. Moreover, in all 52 genomes, operonic genes are more significantly clustered (lower *P*-value) than genes outside operons. **C)** Data in this figure are based on partially or completely randomized *Escherichia coli* K-12 G1655 *(iJO1366)* genomes. We generated four sets of 100 randomized genomes, with randomized orders of i) all genes (black), ii) non-operonic genes (blue), iii) operonic genes (red), and iv) entire operons (yellow, without changing intra-operonic gene orders). The vertical axis shows the extent of clustering of strictly essential genes in the 100 randomized genomes, as indicated by the P-value of Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line indicates the clustering of strictly essential genes in the wild type *Escherichia coli* K-12 G1655 *(iJO1366)* genome, and the red horizontal line shows a minimal threshold of significant clustering (i.e., where Kuiper's test yields P<0.05).

Furthermore, using randomization of metabolic gene orders in the *E. coli K12* genome, we observed that in a partially reshuffled genome where only the positions of operonic genes are randomly reshuffled, the clustering of essential genes is as low as that of completely reshuffled genomes (Figure 2c).In contrast, when we only reshuffled the positions of non-operonic genes, the clustering of essential genes remains similar to that of the wild-type genome. Moreover, when we reshuffled the relative ordering of operons without changing the orders of genes inside any given operon, we also observed a reduction in the clustering of essential metabolic genes (Figure 2c). Similar observations hold for genomes different from that of *E. coli* (figure S17). Thus, we conclude that the metabolic genes in operons that have preserved under deletion pressure and so remained in present-day bacterial genomes contribute to the clustering of essential genes more profoundly than non-operonic metabolic genes. These observations are consistent with the importance of gene deletions in the organization of essential metabolic genes.

Next, we focused on horizontal gene transfer (HGT) to find further signatures of adaptation underlying the clustering of essential genes. If a genome does not experience gene deletions, HGT-acquired genes cannot become essential in the environment in which they have been transferred, because the organism is already viable without the newly transferred genes. However, in the presence of gene deletions, newly and initially non-essential HGT-acquired genes can become essential, for example when other, previously essential genes undergo deletion. Thus, the presence of HGT-acquired essential metabolic genes is a signature of gene deletion events in the genome. Note however that this argument applies only to strictly essential genes (i.e. essential in all environments), not conditionally essential genes, because HGT-aquired genes can become essential in new environments without the need for gene deletions (32).

To search for these signatures of gene deletion, we first identified all HGT-acquired metabolic genes in 43 of the bacterial genomes using the HGTree database (33) (see methods). We then determined what fraction of essential metabolic genes are acquired by HGT. Intriguingly, essential metabolic genes, and in particular strictly essential ones, are more likely to be acquired by HGT than other metabolic genes (figures 3a and S18 and tables S6 and S7). This observation is consistent with a recent experimental study showing that HGT-aquired genes are frequently indispensable (34).

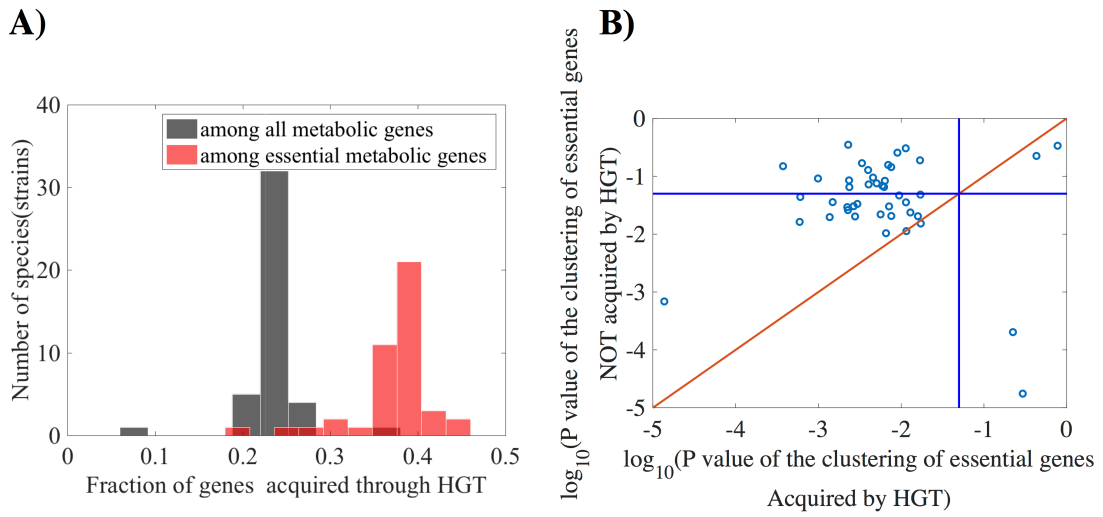



**Figure 3: Horizontal gene transfer and the essential genes. A)** Histogram of the fraction of metabolic genes (black), and the fraction of strictly essential metabolic genes (red) which have been acquired through horizontal gene transfer (HGT) among the 43 species (strains) used in this analysis. We consider a metabolic gene as strictly essential, if its deletion results in losing viability on all the carbon sources on which the wild type metabolism is viable. **B)** In this analysis, we subdivided all strictly essential genes in each metabolic genome into two groups: i) those acquired by horizontal gene transfer (HGT) and ii) those not acquired by horizontal gene transfer. For each of the 43 genomes, and separately for strictly essential genes in each of the two groups, we determined the extent of gene clustering using the P-value generated by Kuiper's test. Each circle in this figure corresponds to a given species or strain. The horizontal and vertical axes show the extent of clustering for genes acquired and not acquired by HGT, respectively. The blue lines correspond to a significance threshold of $P$=0.05 ($-log_{10}0.05$), and the red line is the identity line. Note that horizontally transferred genes show greater clusterin in the vast majority of genomes. Whereas the clustering of horizontally transferred genes is significant at $P$=0.05 in 39 among 43 genomes (90.7%), that of not horizontally transferred genes is significant only in 23 genomes (53.48%). Moreover, horizontally transferred genes are more clustered (higher P-value) than not horizontally transferred genes in 38 genomes (88.37%).

Moreover, significant clustering of HGT-acquired essential genes would provide additional support for gene deletions as an evolutionary force to cluster essential genes. We thus checked whether HGT contributes also to the organization of essential genes in bacterial genomes. To do so we partitioned a genome's set of strictly essential genes into two groups, those acquired by HGT, and those not acquired by HGT. Then, we quantified the clustering of genes in each group separately, using Kuiper's test. In 39 of 43 (90.7%) of our bacterial genomes HGT-acquired essential metabolic genes are significantly clustered. In contrast, non HGT-acquired essential metabolic genes are significantly clustered only in 23 genomes (53.48%; Figure 3b and table S8). Similar observations hold for genes essential for viability on single carbon sources like glucose

(figure S19). In sum, HGT-acquired genes are preferentially essential, and HGT-acquired essential genes are also more clustered than other essential genes, implying that gene deletion contributes to the clustering of essential genes.

So far we have focused on the organization of essential genes, but the organization of non-essential genes may also be important for robustness to large-scale deletions. To find out whether this is so, we first focused on pairs of genes that are individually non-essential but jointly essential, i.e., their simultaneous delection disrupts viability. Such gene pairs are also called synthetic lethals. If two synthetically lethal genes are closely linked in a genome, they are more likely to be deleted together in a tandem deletion. In contrast, if they are far away from each other, the likelihood that both of them are deleted in the same tandem deletion is much lower. Thus, synthetically lethal genes that are further apart than expected by chance alone could lead to increasing robustness to tandem gene deletion. We refer to such synthetically lethal genes as being in repulsion.

To find out whether synthetically lethal genes are in repulsion, we created pairwise deletions of all non-essential metabolic genes in all 55 prokaryotic genomes, and determined their viability. In this analysis, we made a distinction between two types of synthetically lethal genes. The first comprises *strictly synthetically lethal* gene pairs, whose joint deletion is lethal in all carbon source environments we consider. The second comprises conditionally synthetic lethal gene pairs, whose deletion is lethal in at least one but not all environments. We determined the distance between two strictly synthetically lethal metabolic genes as the number of metabolic genes that lie between them. In the majority of genomes (41 out of 55; 74.54%), at least 50 genes lie between all strictly synthetic lethal gene pairs (table S13), and the paucity of strictly synthetically lethal gene pairs with a distance below 50 is statistically significant (table S14; Fisher's exact test). This repulsion is also evident from a circos plot of the *E. coli* genome (figure 4a), and it disappears after random genome shuffling (Figure 4b). No short-range synthetic lethal interactions exist in the *E. coli* genome, but in the randomized genome, such interactions are abundant (Figures 4b and 4c). Similar patterns exist in other species (Figures S20 and S21). The same does not hold for conditionally lethal gene pairs (tables S15 and S16) and some bacterial species with small metabolic genome sizes (Figure S22), which suggests that this repulsion might be the result of long DNA insertions during bacterial genome growth.

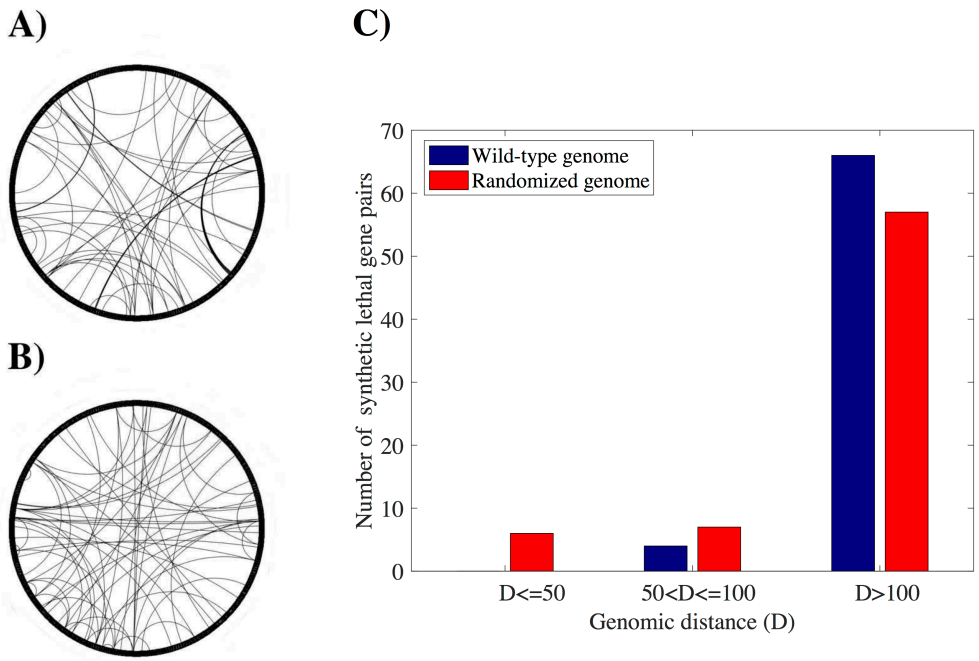



**Figure 4: Repulsion of synthetic lethal genes. A)** Circos plot of the *Escherichia coli* K-12 G1655 *(iJO1366)* genome, in which metabolic genes are arranged according to their order in the genome. An arc connects two genes if they form an unconditionally synthetic lethal pair. **B)** Same as A, but for randomized gene order. Note the many more short-ranged synthetic lethality interactions after gene order randomization. **C)** Barplot of the genomic distance (in number of intervening genes) between unconditionally synthetic lethal metabolic gene pairs in the wild-type (blue) and randomized (yellow) *Escherichia coli* K-12 G1655 *(iJO1366)* genome. Note the lack of short-distance synthetic lethal pairs with fewer than 50 intervening genes in the wild type genome.

The role of nonessential genes in robustness to gene deletions may not be restricted to pairs of such genes, but could be extended to three or more genes that are individually nonessential but jointly essential. Beyond two genes the number of possible combinations of such synthetically lethal *n*-tuples of genes becomes too large for exhaustive analysis. However, if such genes are in repulsion, genomes might be enriched in long clusters of genes that harbor no essential genes, and whose joint deletion is not lethal. We showed that such gene clusters do indeed exist (see text S5, figure S23, and tables S17 and S18). However, similar to what we observed for clusters of essential genes, genome randomization forces such as translocation and inversion can erode non-essential gene clusters (figures S24 and S25). Thus, the maintenance of these non-essential clusters in the genome would also require similar adaptive processes.

In summary, we have shown that the ordering of metabolic genes in bacterial genomes provides phenotypic robustness against deleterious effects of large-scale gene deletions. This robustness can endow bacterial populations with the flexibility to survive large-scale gene deletion events, which could potentially help them adapt to new environments (particularly in pathogenic species) (35–38). Underlying this excess robustness is a non-random distribution of both essential and non-essential metabolic genes, which is manifested as clustering of essential genes and repulsion of synthetic lethal genes. Although a genome growth process starting from minimal genomes shows that clustering of essential genes could in principle have non-adaptive origins, the significant contribution of HGT and operons to this organization raises the possibility that the emergence of this genomic ordering from randomness is an adaptation to frequently-occuring large-scale gene deletions.

## 3.5. Methods

**Bacterial genome-scale metabolic networks:** We used 55 reconstructed bacterial genome-scale metabolic networks from the BiGG database (14), which provides comprehensive information about biochemical reactions, metabolites, metabolic genes, and gene-reaction association rules for each bacterial species. We ordered the genes in each species based on their genomic location, as obtained from the RefSeq microbial genome database (39). We used the R-package Sybil (40) to parse the BiGG models.

**Phenotype prediction from genomic information:** We focus our analyses on a qualitative definition of metabolic phenotypes, that is, on whether a given metabolism is *viable* or *inviable* in a given minimal chemical environment (medium) that contains only a single carbon source. More specifically, we consider a genotype viable if it can produce all essential biomass precursors from the resources in this medium. We use Flux Balance Analysis (FBA, See text S6) to predict viability (12).

Among the 55 bacterial metabolisms we study, we identified 137 unique carbon-containing metabolites that occur in the metabolism of all species. Thus, we considered 137 minimal growth environments that were distinguished by these carbon sources. Each of these environments included one carbon source, as well as oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe2+ and Fe3+), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese,

and zinc. In other words, we varied the carbon source while keeping all other nutrients constant. None of the 55 species we studied were viable on 34 of these 137 minimal environments, so we excluded the corresponding carbon sources from this study, and performed our analysis with the remaining 103 minimal environments, whose carbon sources are listed in table S1.

To systematically examine viability after deleting a metabolic gene or a set of metabolic genes, we used gene-reaction association rules for each species obtained from the BiGG database(14). Based on these rules, we translated metabolic gene deletions into deleted reactions. For more than 10% of reactions, genes and reactions do not show a one-to-one association. Some reactions are catalyzed by one or more enzymatic complexes, which may be encoded by more than one metabolic gene. In this case, deletion of a single gene whose product participates in a given enzymatic complex is enough to inactivate the complex (i.e., a Boolean AND function of gene presence/absence determines whether a reaction can be catalyzed). Other reactions can be catalyzed independently by multiple enzymatic complexes. In this case, all complexes need to be inactivated by deletion of individual genes to eliminate a reaction from the metabolic network (corresponding to a Boolean OR function of complex activity/inactivity). Finally, some gene products may participate in multiple enzymatic reactions, such that deletion of a single gene would eliminate multiple reactions. We took these associations into account when translating gene deletions into reaction deletions. After any one such deletion, we determined with FBA whether the resulting metabolic network is still viable in any one environment.

**Quantification of robustness to multiple gene deletions:** To quantify the robustness of a given genome (metabolism) with $n$ metabolic genes to "tandem deletions" of length $l$ genes in a given environment (carbon source), we considered all possible ($n$) deletional variants in each of which $l$ consecutive metabolic genes are deleted. For each deletional variant, we determined the reactions to be deleted from the wild-type metabolic network, based on the gene-reaction association rules (14). Subsequently, we determined metabolic viability of each variant by FBA, and quantified the robustness to tandem deletion as the fraction of deletional variants that retain viability on the given carbon source.

To quantify the robustness of a given genome (metabolism) with $n$ metabolic genes to a "random deletion" of length $l$, in a given environment, we generated the same number

*n* of deletional variants as for tandem deletions. In each of these variants *l* randomly chosen metabolic genes in the genome are deleted (irrespective of their genomic location). We quantified robustness to random gene deletion with the same procedure described above, as the fraction of random deletional variants that retain viability on the carbon source.

**Quantification of gene essentiality:** To determine whether a metabolic gene is essential for viability on a given carbon source, we removed the corresponding reaction(s) from the wild-type metabolic network, and determined viability using FBA. For each bacterial genome, we determined the essentiality of every metabolic gene in every environment on which the wild-type metabolism is viable. We consider a metabolic gene as *"strictly essential"* in a given genome, if its deletion results in losing viability on all carbon sources on which the wild-type metabolism is viable, and we consider a metabolic gene as *"conditionally essential"* if its deletion abolishes viability on at least one carbon source. Note that strictly essential genes are a subset of conditionally essential genes.

Likewise, we call a metabolic gene *"strictly non-essential"* if its deletion does not abolish viability on any carbon source, and we indicate a metabolic gene as *"conditionally non-essential"* if its deletion does not abolish viability on at least one carbon source. Strictly non-essential genes are a subset of conditionally non-essential genes.

**Quantification of the clustering of essential genes in a given genome:** We used Kuiper's test (23) to assess whether the distribution of essential genes in a given genome is uniform or not. This test is closely related to the Kolmogorov-Smirnov test (K-S test), which computes the discrepancy statistics $D^+$ and $D^-$ that represent the absolute sizes of the most positive and most negative differences between two cumulative probability distribution functions that are being compared. Because, the K-S test is not invariant under cyclic transformations, it is not useful to detect clusters of genes distributed in a circular bacterial genome. Kuiper's test allows cyclic transformations while taking advantage of the $D^+$ and $D^-$ test statistics.

**Generation of minimal metabolic genome:** We define a minimal metabolic genome as a set of metabolic genes of a given species that are all necessary to produce essential biomass precursors from external nutrients available in a given environment. To create

a minimal genome, one needs to delete the non-essential genes step by step until no non-essential metabolic genes remain. Note that the size of minimal genome may be larger than the number of essential genes in a wild-type (full-sized) genome, because after deleting a given gene, some previously non-essential genes may become essential. Moreover, genome size and gene identities in a minimal genome depend on the order of gene deletions that occur during genome reduction(28).

To generate a minimal genome from a given full-size genome, we apply a previously established stepwise stochastic algorithm(28). In each step, we remove a randomly chosen metabolic gene from the genome and determine the viability of the resulting metabolism in the given environment. If the metabolism is still viable (i.e. it can produce all biomass precursors), we accept the deletion and remove the gene from the genome; otherwise, the gene is restored to the genome. This procedure is repeated until no further genes can be deleted, that is, until all remaining metabolic genes are essential for survival in the given environment. We applied this procedure to three different genomes, namely to *Escherichia coli* K-12 G1655 *(iJO1366), Bacillus subtilis,* and *Salmonella enterica*, using glucose or acetate as carbon sources. From each genome and on each carbon source, we generated 100 different minimal genomes.

**Identification of horizontally transferred metabolic genes:** We used the HGTree database (33) to identify the metabolic genes that any one genome has likely obtained through horizontal gene transfer. In this database, horizontally transferred genes are predicted based on a tree reconciliation method, which reconstructs approximate maximum likelihood phylogenetic trees for each orthologous gene and corresponding 16S rRNA reference species sets, and then reconciles the two trees using maximum framework. Because 43 of the 55 bacterial species that we considered were included in this database, we focused this part of our analysis on 43 bacterial genomes.

**Identification of operons in bacterial genomes:** To identify operons, we used the DOOR database(18), which is a comprehensive database for prokaryotic operon information to identify metabolic genes that belong to an operon. It predicts operons based on a computational method(19) that was ranked first in an independent assessment of 14 operon prediction methods(41). For genomes with many experimentally validated operons, this method predicts operons based on a decision-tree based classifier that uses both genome-specific features such as conserved gene neighborhood, phylogenetic profiles and intergenic distances, and general features such

as the length ratio between a pair of adjacent genes, Gene Ontology (GO)-based functional similarity between adjacent genes and the frequency of a specific DNA motif in the intergenic region. In contrast, for genomes with only limited experimental data on operons, the program applies a logistic function-based classifier using solely general genome features. The DOOR database contained operon information for 52 of our 55 bacterial genomes.

**Identification of pairs of synthetic lethal genes:** For any given genome (metabolism), we identified all genes that are non-essential for viability in a given environment. Then, we examined all pairs of non-essential genes to determine whether simultaneous deletion of these genes is lethal. If yes, we consider the pair of genes as a synthetic lethal pair in this environment. We call a pair of genes that are synthetic lethal in *all* environments on which a wild-type metabolism is viable, *unconditionally* synthetic lethal genes. Conversely, we call pairs of genes that are synthetically lethal in some but not all environments *conditionally* synthetically lethal.

**Identification of non-essential clusters of non-essential metabolic genes:** Any two successive strictly essential metabolic genes are either adjacent in the metabolic genome or non-adjacent, i.e., separated by one, two, or a larger cluster of non-essential metabolic genes. The non-essential genes belonging to such a cluster are at least conditionally non-essential, meaning that they are non-essential on at least one carbon source, but they are not necessarily strictly non-essential. Thus, we call a cluster of non-essential metabolic genes intervening between two successive strictly essential metabolic genes a "*cluster of conditionally non-essential metabolic genes"*. After identifying all clusters of conditionally non-essential metabolic genes, we aimed to determine whether each such cluster is essential for viability on the carbon sources we consider. To do so, we deleted all metabolic genes in a given cluster, translated the gene deletions into reaction deletions, and used FBA to determine the resulting metabolism's viability on the set of carbon sources on which the wild-type metabolism (before deletion) was viable. If the deletion did not abolish viability on any of these carbon sources, we called such a cluster a "*strictly non-essential cluster of conditionally non-essential metabolic genes*". Moreover, if the deletion did not abolish viability on at least one carbon source on which the wild-type metabolism was viable, we called the cluster a "*conditionally non-essential cluster of conditionally non-essential metabolic genes*". Note that the set of strictly non-essential clusters (of

conditionally non-essential metabolic genes) is a subset of the set of conditionally non-essential clusters (of conditionally non-essential metabolic genes).

We can apply a similar procedure for any cluster of *strictly non-essential* metabolic genes intervening between two successive (but not adjacent) conditionally essential metabolic genes. In this way, we can identify *strictly non-essential clusters of strictly non-essential metabolic genes* and *conditionally non-essential clusters of strictly non-essential metabolic genes*. Note that again the set of strictly non-essential clusters (of strictly non-essential metabolic genes) are a subset of the set of conditionally non-essential clusters (of strictly non-essential metabolic genes). However, the set of clusters of strictly non-essential metabolic genes are not a subset of the set of the clusters of conditionally non-essential metabolic genes.

## 3.6. References


1. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3(9):711–21.
2. Lynch M (2006) Streamlining and Simplification of Microbial Genome Architecture. *Annu Rev Microbiol* 60(1):327–349.
3. McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 10(1):13.
4. Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108(5):583–6.
5. Wolf YI, Koonin E V. (2013) Genome reduction as the dominant mode of evolution. *BioEssays* 35(9):829–837.
6. Albalat R, Cañestro C (2016) Evolution by gene loss. *Nat Rev Genet* 17(7):379–391.
7. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17(10):589–96.
8. Kunin V, Ouzounis CA (2003) The Balance of Driving Forces During Genome Evolution in Prokaryotes. *Genome Res* 13(7):1589–1594.
9. Nilsson AI, et al. (2005) Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci U S A* 102(34):12112–6.
10. Lee M-C, Marx CJ, Lenski R, Sivam D, Lidstrom M (2012) Repeated, Selection-Driven Genome Reduction of Accessory Genes in Experimental Populations. *PLoS Genet* 8(5):e1002651.
11. Koskiniemi S, Sun S, Berg OG, Andersson DI, Boxer D (2012) Selection-Driven Gene Loss in Bacteria. *PLoS Genet* 8(6):e1002787.

12. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245–8.

13. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* 9:661.

14. King ZA, et al. (2015) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*:gkv1049-.

15. Lawrence JG, Roth JR (1996) Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. *Genetics* 143(4):1843–1860.

16. Pál C, Hurst LD (2004) Evidence against the selfish operon theory. *Trends Genet* 20(6):232–234.

17. Price MN, Huang KH, Arkin AP, Alm EJ (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* 15(6):809–19.

18. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res* 37(Database):D459–D463.

19. Dam P, Olman V, Harris K, Su Z, Xu Y (2006) Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 35(1):288–298.

20. Rocha EPC, Danchin A (2003) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res* 31(22):6570–7.

21. Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? *Mol Biol Evol* 22(11):2147–56.

22. Fang G, Rocha EPC, Danchin A (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics* 9:4.

23. Kuiper NH (1960) Tests concerning random points on a circle. *Indag Math* 63:38–47.

24. Mushegian A (1999) The minimal genome concept. *Curr Opin Genet Dev* 9(6):709–14.

25. Glass JI, et al. (2006) Essential genes of a minimal bacterium. *Proc Natl Acad Sci U S A* 103(2):425–30.

26. Posfai G, et al. (2006) Emergent Properties of Reduced-Genome Escherichia coli. *Science (80- )* 312(5776):1044–1046.

27. Hutchison CA, et al. (2016) Design and synthesis of a minimal bacterial genome. *Science (80- )* 351(6280):aad6253-aad6253.

28. Pál C, et al. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440(7084):667–670.

29. Hill CW, Gray JA (1988) Effects of chromosomal inversion on cell fitness in Escherichia coli K-12. *Genetics* 119(4):771–8.

30. Segall A, Mahan MJ, Roth JR (1988) Rearrangement of the bacterial chromosome: forbidden inversions. *Science* 241(4871):1314–8.

31. Itoh T, Takemoto K, Mori H, Gojobori T (1999) Evolutionary instability of

operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* 16(3):332–46.

32. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37(12):1372–5.

33. Jeong H, et al. (2016) HGTree: database of horizontally transferred genes determined by tree reconciliation. *Nucleic Acids Res* 44(D1):D610-9.

34. Karcagi I, et al. (2016) Indispensability of Horizontally Transferred Genes and Its Impact on Bacterial Genome Streamlining. *Mol Biol Evol* 33(5):1257–1269.

35. Hottes AK, et al. (2013) Bacterial Adaptation through Loss of Function. *PLoS Genet* 9(7):e1003617.

36. Sokurenko E V, Hasty DL, Dykhuizen DE (1999) Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol* 7(5):191–5.

37. Maurelli AT, Fernández RE, Bloch CA, Rode CK, Fasano A (1998) "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli. *Proc Natl Acad Sci U S A* 95(7):3943–8.

38. Moore RA, et al. (2004) Contribution of Gene Loss to the Pathogenic Evolution of Burkholderia pseudomallei and Burkholderia mallei. *Infect Immun* 72(7):4172–4187.

39. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* 42(Database issue):D553-9.

40. Gelius-Dietrich G, Desouki AA, Fritzemeier CJ, Lercher MJ (2013) Sybil--efficient constraint-based modelling in R. *BMC Syst Biol* 7:125.

41. Brouwer RWW, Kuipers OP, van Hijum SAFT (2008) The relative value of operon predictions. *Brief Bioinform* 9(5):367–375.

## 3.7. Supplementary Information

### Text S1: The selfish operon hypothesis cannot explain the excess robustness to tandem gene deletions

In this analysis, we identified all operons and their corresponding metabolic genes in 52 bacterial genomes using the DOOR database(1). For each genome, we generated all operon deletion variants, i.e., we selected one of the operons and deleted all metabolic genes belonging to it, and repeated this procedure for all operons. We then used flux balance analysis to determine the viability of each operon deletion variant on 102 distinct carbon sources. We quantified the robustness of a given genome to operon deletion ($R_{Operonic}$) as the fraction of operon deletion variants that retain viability in a given environment or environments. To compare this robustness with the average robustness to tandem deletions of the same length for a given bacterial genome, we measured the weighted average of robustness to tandem deletions ($\bar{R}_{tandem}$), which we define as $\bar{R}_{tandem} = \sum_{n=2}^{20} w_n R^n_{tandem}$, where $w_n$ is the fraction of operons with *n* metabolic genes in the analyzed genome, and $R^n_{tandem}$ is the genome's robustness to tandem deletions of *n* metabolic genes. Note that *n* varies between 2 and 20, because the smallest and largest operons in our study genomes have this respective number of genes. This weighted average ensures that the average length of tandem deletions that enter the calculation is the same as the average length of operons in any given genome. Figure S7 and table S2 show that robustness to operon deletion is slightly higher than the average robustness to tandem deletion, but the difference is not as dramatic as that between robustness to tandem deletion and random deletion. Thus, the dispensability of operons as implied by selfish operon theory cannot fully explain the excess robustness to tandem deletions.

### Text S2: Essential metabolic genes are clustered in bacterial genomes

We hypothesized that the clustering of essential genes can increase the robustness of the genome to simultaneous deletion of multiple successive (*tandem*) genes. In contrast, it should not impact robustness to simultaneous deletions of multiple genes, chosen at random, regardless of their genomic location. If so, the clustering of essential genes might to a large extent explain the excess robustness to tandem

deletions. To find out whether this is the case, we first identified all essential metabolic genes in each of our 55 study genomes and in each of the 102 minimal environments (Table S3). We then used the test statistic of Kuiper's test (Methods) to calculate a measure of clustering, i.e., the extent to which the distribution of essential genes differs from a uniform distribution. In this analysis we distinguished two types of essential genes: $i$) strictly essential metabolic genes, which are essential on all carbon sources, and $ii$) conditionally essential metabolic genes, which are essential on at least one carbon source (see methods). In the vast majority of the species, both types of essential genes are significantly clustered in the genome (Tables S4 and S5). What is more, we observed a positive correlation between the degree of essential gene clustering and robustness to tandem gene deletions (deletion length 5: Pearson's *r=-0.26*, and *P=0.05;* length 10; *r=0.35*, and *P =0.009*)), but no significant correlation with robustness to random deletions (of length 5 (Pearson's *r=-0.04*, and *P-value=0.77*) and of length 10 (Pearson's *r=-0.01*, and *P-value=0.95*)). This confirms the importance of essential gene clustering for robustness to tandem gene deletions.

**Text S3: Passive emergence of the clusters of essential genes**

We hypothesized that an evolutionary genome expansion scenario could explain essential gene clustering in present-day bacterial genomes. To validate this hypothesis, we started from minimal genomes(2), that is, sets of metabolic genes from which not a single gene can be removed without abolishing viability on a specific carbon source (See methods in the main text). Specifically, we started with 100 distinct minimal metabolic genomes, which we had derived from the *E. coli K12* wild type genome through stepwise elimination of genes that are nonessential for viability on glucose as the sole carbon source. Starting from one of these minimal genomes, we then chose randomly (with a uniform distribution) between 1 and 5 not necessarily contiguous metabolic genes from the wild type *E. coli K12* genome that are not already included in the minimal genome. We then inserted the selected genes as a contiguous set of genes into a randomly chosen position in the genome. This procedure implies that during an insertion event on average 3 genes are added to the genome. This choice of including multiple genes in an insertion event is motivated based on ample empirical evidence in favor of co-acquisition and co-insertion of multiple genes into bacterial genomes, for example during horizontal gene transfer events(3–7). We repeated this insertion process, taking care to only choose genes for

insertion that were not already present in the growing recipient genome, until our genome had reached the size of the wild type *E. coli K12* genome. We note that the gene content of the resulting genome is identical to that of the wild type genome, but its gene order is not. We repeated this procedure for all 100 starting minimal genomes. We observed that the robustness of the resulting genomes to tandem gene deletion (figure S9a (blue boxes)) is considerably higher than that of a randomly reshuffled *E. coli K-12* genome (figure S9a (black boxes)), and comparable to that of the *E. coli K-12* wild-type genome (figure S9a (blue horizontal line)). Because the final genomes produced by these simulations have the same set of genes as that of *E. coli K-12*, the fraction of essential genes is exactly the same as that of *E. coli K-1*2 (Figure S9b). A difference in this fraction to the *E. coli K-1*2 can thus not possibly be responsible for the increase in robustness (figure S9b). In the majority of the resulting genomes, essential genes are also significantly clustered (figure S9c). Thus, essential genes can passively get clustered in the genome as a byproduct of increasing genomic complexity.

We then examined additional evolutionary forces, such as gene deletion and duplication, which might further enhance robustness to tandem deletions. More specifically, we compared four different kinds of genome-altering changes to expand each of 100 minimal genomes derived from the *E. coli K-12* genome and viable on glucose to a size that is identical to that of the *E. coli* K-12 genome. That is, we expanded genomes through: $i$) insertion events alone (as just described), $ii$) insertion and deletion events, $iii$) insertion and duplication events, and $iv$) insertion, duplication, and deletion events, as described below:

*Insertion + deletion*: We started with 100 distinct minimal genomes derived from the *E. coli K-12* genome and viable on glucose. For each of these genomes, we performed the following iterative procedure in each step of which we either $i$) (with 75% probability) randomly and with a uniform distribution chose between 1 to 5 (not necessarily contiguous) metabolic genes from the set of genes that were present in the full-size genome, but absent from the minimal genome, and inserted them as a contiguous gene cluster at a randomly chosen position in the recipient genome, or $ii$) (with 25% probability) deleted a randomly chosen non-essential cluster of genes in the growing genome. We iterated this procedure until the growing genome had reached a size equal to that of the focal species. Note that this procedure also avoids

duplication of genes in the growing genome, and ensures that all genes in the full-size wild type genome will be included in the final genomes. The higher insertion than deletion probability ensures that genome size grows over time. Note that including deletions implies that more steps are needed to reach the final genome size.

*Insertion + duplication*: We started with 100 distinct minimal genomes derived from the *E. coli K-12* genome and viable on glucose. For each of these genomes, we performed the following iterative procedure in each step of which we either $i$) (with 75% probability) randomly and with a uniform distribution chose between 1 to 5 (not necessarily contiguous) metabolic genes from the set of genes that were present in the full-size genome, but absent from the minimal genome, and inserted them as a contiguous gene cluster at a randomly chosen position in the recipient genome, or $ii$) (with 25% probability) we duplicated a given number of genes chosen at random. We iterated this procedure until the growing genome had reached a size equal to that of the focal species. In these simulations, gene duplications contributed to 25% of the added genes, which implies that the final genome will not contain some of the genes in the wild-type genome of the focal species.

*Insertion + duplication + deletion*: We started with 100 distinct minimal genomes derived from the *E. coli K-12* genome and viable on glucose. For each of these genomes, we performed the following iterative procedure in each step of which we either $i$) (with 50% probability) randomly and uniformly chose between 1 to 5 (not necessarily contiguous) metabolic genes from the set of genes that were present in the full-size genome, but absent from the minimal genome, and inserted them as a contiguous gene cluster at a randomly chosen position in the recipient genome, or $ii$) (with 25% probability) duplicated a given number of genes chosen at random, or $iii$) (with 25% probability) deleted a randomly chosen non-essential cluster of genes in the growing genome. We iterated this procedure until the growing genome had reached a size equal to that of the focal species.

At the end of each simulation, and for each of the 100 final full-sized genomes, we quantified metabolic robustness to tandem deletions of length 5, identified essential genes in each genome, and studied whether these genes are clustered.

We observed that most of the genomes subject to both insertion and deletion reached higher final robustness to tandem deletions than for the other three scenarios, and

even higher robustness than wild-type *E. coli* (figure S9a (cyan boxes)). Thus, exposing our simulated genomes to tandem deletions leads to higher robustness to such deletions. Including duplications slightly lowers robustness compared to the insertion-only scenario. To understand why, consider the following. In the absence of gene duplication, the final genome contains the same genes as that of *E. coli K12*, albeit with different order. In the presence of duplication, however, some *E. coli K12* genes may have multiple copies in the final genome, whereas others may be missing. The fraction of essential genes is the same as that of *E. coli K12* in the absence of duplication, but it is smaller in the presence of duplication, because duplication of an essential gene results in two individually non-essential genes (figure S9b). Furthermore, in the presence of gene duplication, essential genes do not become significantly clustered, whereas under insertion and deletion alone, essential genes in the second scenario are more strongly clustered than *E. coli K12* (figure S9c). Because gene duplication disrupts clusters of essential genes, it does not increase robustness to tandem gene deletion despite lowering the fraction of individually essential genes.

To check whether the ordering of the metabolic genes in the initial minimal genomes is important for the patterns we observed, we repeated the above procedure starting with minimal genomes in which the ordering of the metabolic genes are randomly reshuffled. Using this approach we observed the same patterns (figure S10). Moreover, our observations remain the same when we use acetate instead of glucose as the sole carbon source in the minimal medium (figure S11). Finally, we get similar results when we repeat this procedure for other bacterial species (figure S12).

In sum, we conclude that clusters of non-essential genes observed in wild-type genomes might originate non-adaptively. In other words, they can emerge passively through sequential insertion of new genes into a minimal genome. Moreover, exposure to gene deletion enhances clustering of essential genes, but duplication can disrupt clusters of essential genes.

**Text S4: Genome rearrangement can disrupt essential gene clusters**

To study the effects of genome rearrangements on the organization of bacterial genomes, we used computer simulations that start out with 100 copies of wild type genomes of *E. coli K12*, and 100 copies of the wild type genomes of *Salmonella*

*enterica*. We subjected each genome to 1000 independent stochastic genomic rearrangement events, studying three distinct scenarios of genome rearrangement events: *i)* translocation alone, *ii)* inversion alone, and *iii)* translocation + inversion. In any translocation step, a given number of consecutive metabolic genes (between 2 to 10) is randomly chosen and is translocated to a randomly chosen position in the genome, while in an inversion step, the genes are left in place but the relative ordering of the genes is reversed. After each rearrangement step, we quantified the robustness of each of the 100 genomes to tandem gene deletions, as determined by the fraction of encoded metabolisms viable on glucose and acetate. Moreover, we identified essential genes in each genome and studied whether these genes are clustered.

In all three scenarios, genome rearrangement events gradually reduce, for both species, robustness to tandem gene deletions on glucose (Figure S13a-c). Inversion alone reduces robustness to a lesser extent than the other two scenarios. Moreover, translocation reduces the clustering of essential genes, but inversion alone does not affect this clustering (Figure S13d-f). The same holds on acetate instead of glucose (Figure S14). Further analysis shows that inversion reduces robustness only by shrinking clusters of non-essential genes, whereas translocation both disrupts essential gene clusters and shrinks clusters of non-essential genes (see text S5 and figures S24 and S25).

**Text S5: Long non-essential clusters of non-essential genes**

We observed that bacterial genomes harbor long clusters of non-essential metabolic genes that are not interrupted by any essential genes. If simultaneous deletion of all these non-essential genes does not abolish viability, then this arrangement in itself increases the robustness of bacterial genomes to tandem deletions. We distinguish between two different types of non-essential metabolic genes: *i)* strictly non-essential genes that are not essential on any carbon sources, and *ii)* conditionally non-essential genes that are not essential on at least one carbon source. By definition, a cluster of strictly non-essential genes intervenes between two successive but non-adjacent conditionally essential genes, and a cluster of conditionally non-essential genes intervenes between two successive but non-adjacent strictly essential genes.

We observed that both strictly and conditionally non-essential genes are organized into fewer but longer clusters in wild-type genomes than in randomized genomes

(Figure S23). For example, in *Escherichia coli* K-12 G1655 *(iJO1366)*, we observed a long cluster of 32 consecutive strictly non-essential metabolic genes and a cluster of 65 consecutive conditionally non-essential metabolic genes. We observed that simultaneous deletion of all the genes in a given cluster of strictly non-essential metabolic genes does not abolish viability on any carbon sources in more than 95% of the clusters, and it does not abolish viability on at least one carbon source in more than 99% of the clusters for the vast majority of the genomes (50 out of 55, that is 90.9%) (See table S17). Moreover, in around 90% of the clusters of conditionally non-essential metabolic genes, simultaneous deletion of all the genes belonging to the cluster does not abolish viability on at least one carbon source, and in around 50% of these clusters in all genomes, it does not abolish viability on any carbon source (See table S18). This carbon-source dependent viability loss caused by deletion of the clusters of non-essential genes can explain the variability among different carbon sources that we observed (Figures S5 and S6).

Finally, we showed that genome rearrangements like translocations and inversions shrink the length of non-essential gene clusters (figures S24 and S25). In figures S13 and S14 we observed that inversions alone causes a gradual decline in robustness to tandem gene deletions without any reduction in the clustering of essential genes. Thus, it is the shortening of the clusters of non-essential genes caused by gene inversions that reduces robustness to tandem deletions (Figures S24 and S25).

**Text S6: Flux balance Analysis**

Flux balance analysis (FBA) is a widely used computational method for the quantitative analysis and modeling of metabolic networks (8). FBA predicts the metabolic flux through each reaction in a given metabolic network using the stoichiometric coefficients of metabolites participating in the network's reactions. Stoichiometric coefficients are stored in a stoichiometric matrix *S*, which is of dimension $m \times n$, where $m$ and $n$, respectively, denote the number of metabolites and the number of reactions in a metabolic network. FBA constrains the flux through each reaction based on the assumption that the metabolic network is in a steady state in which metabolite concentrations do not change, i.e., $Sv = 0$, where $v$ is the vector of metabolic fluxes $v_i$ through reaction $i$. The solutions of the equation $Sv = 0$, that is, the null space of the matrix S, comprises all flux vectors that are allowable in steady state. The null space can be further constrained by physicochemical information

regarding the maximally and minimally possible fluxes through each reaction. FBA relies on linear programming to identify those allowable flux vector(s) that maximize an objective function *Z*. This task can be formulated as finding a flux vector $v^*$ with the property

$$v^* = max_v\, Z(v) = max_v \{ c^T v \mid Sv = 0,\ a \leq v \leq b \},$$

where the vector $c$ contains a set of scalar coefficients representing the maximization criterion, and each entry $a_i$ and $b_i$ of vectors $a$ and $b$, respectively, indicate the minimally and maximally possible flux through reaction $i$. The vector $c$ represents the proportions of each small biomass molecule in a cell's biomass. The quantity $v^*$ maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass (8). Here we use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and we consider a metabolic network viable if it can produce all essential biomass precursors. In a free-living bacterium like *E. coli,* there are approximately 60 such molecules including 20 amino acids, DNA, and RNA precursors, lipids and cofactors. We used the biomass composition of the *E. coli* metabolic model iAF1260 to define the vector $c$ (9). Moreover, we used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problem of FBA. The major limitation of FBA is that it neglects regulatory constraints that can arise through suboptimal expression or regulation of enzymes. Newly horizontally transferred genes cannot easily establish regulatory interactions with their host genes, and it may thus take considerable adaptive evolution until they become expressed at a maximal or optimal level (10). Such regulatory constraints would be especially important if we focused on quantitative predictions of biomass growth (11). However, we use FBA solely for qualitative predictions of viability. This focus on qualitative phenotypes is biologically sensible. The reason is that many organisms grow slowly in their native environment (12, 13), implying that regulation for maximal biomass production is far from universal. Moreover, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments (11, 14, 15).

## Supplementary References:


1. Mao F, Dam P, Chou J, Olman V, Xu Y (2009) DOOR: a database for prokaryotic operons. *Nucleic Acids Res* 37(Database):D459–D463.
2. Pál C, et al. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440(7084):667–670.
3. Chan CX, Beiko RG, Darling AE, Ragan MA (2010) Lateral Transfer of Genes and Gene Fragments in Prokaryotes. *Genome Biol Evol* 1(0):429–438.
4. Dilthey A, Lercher MJ (2015) Horizontally transferred genes cluster spatially and metabolically. *Biol Direct* 10:72.
5. Igarashi N, et al. (2001) Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J Mol Evol* 52(4):333–41.
6. Omelchenko M V, Makarova KS, Wolf YI, Rogozin IB, Koonin E V (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol* 4(9):R55.
7. Lin CH, Bourque G, Tan P (2008) A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol Biol Evol* 25(3):549–58.
8. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nat Biotechnol* 28(3):245–8.
9. Feist AM, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.
10. Lercher MJ, Pál C (2008) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol Biol Evol* 25(3):559–67.
11. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420(6912):186–9.
12. Vieira-Silva S, Rocha EPC (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet* 6(1):e1000808.
13. Kirschner D, Marino S (2005) Mycobacterium tuberculosis as viewed through a computer. *Trends Microbiol* 13(5):206–11.
14. Fong SS, Palsson BØ (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet* 36(10):1056–8.
15. Fong SS, Marciniak JY, Palsson BO (2003) Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J Bacteriol* 185(21):6400–6408.

## Supplementary Figures

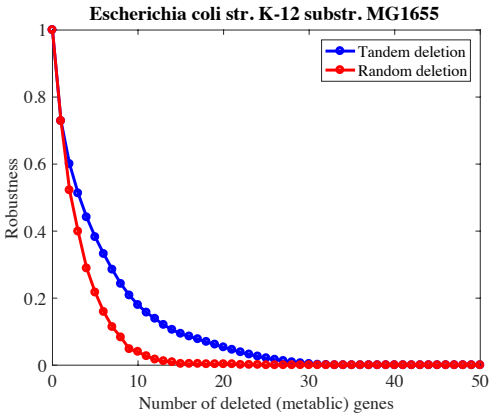



**Figure S1: Robustness to tandem deletion versus random deletion (strict phenotype definition).** **A)** The vertical axis shows the robustness of *Escherichia coli* K-12 G1655 *(iJO1366)* to tandem (blue) and random (red) deletion of metabolic genes, averaged over all deletional variants we examined, as a function of the number of deleted genes (horizontal axis). In this analysis, robustness is defined as the fraction of deletional variants that retain viability on all 97 carbon sources on which the wild type *Escherichia coli* K-12 G1655 *(iJO1366)* is viable. Interpolation between data points is linear and is displayed as a visual guide.

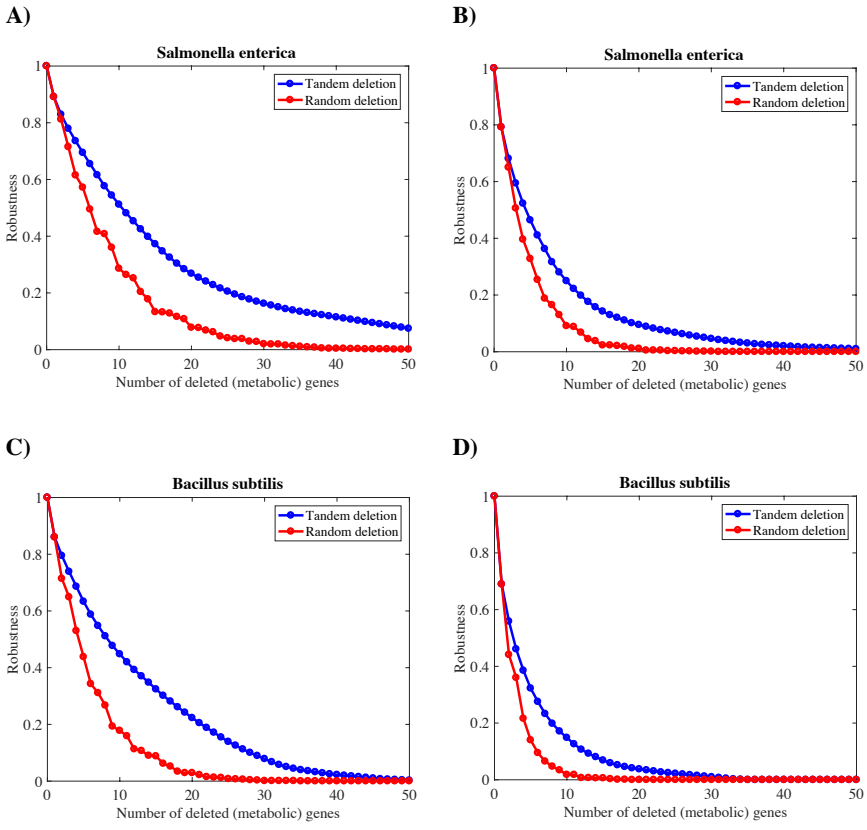



**Figure S2: Robustness to tandem deletion versus random deletion (using alternative bacterial genomes).** The vertical axes show the robustness of *Salmonella enterica* (panels **A** and **B**) and *Bacillus subtillis* (panels **C** and **D**) to tandem (blue) and random (red) deletion of metabolic genes, averaged over all deletional variants we examined, as a function of the number of deleted genes (horizontal axis). Robustness in panels A and C is defined conditionally, i.e., as the fraction of deletional variants that retain viability on at least one carbon source, while in panels B and D it is defined strictly, i.e., as the

fraction of deletional variants that retain viability on all carbon sources on which the wild type metabolism is viable. Interpolation between data points is linear and is displayed as a visual guide.

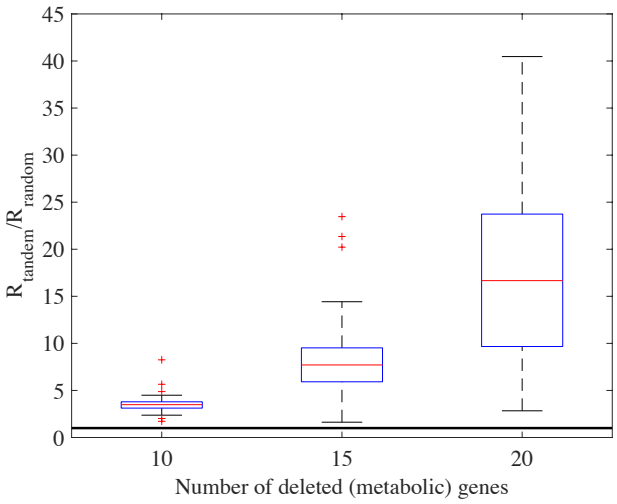



**Figure S3: Excess robustness to tandem deletion (strict phenotype definition)**. The vertical axis shows the excess robustness to tandem deletion, defined as the ratio of robustness to tandem deletion and robustness to random deletion ($R_{tandem}/R_{random}$), for all 55 bacterial genomes, as a function of the number of deleted genes (horizontal axis). In this analysis, robustness is defined as the fraction of deletional variants that retain viability on all carbon sources on which the wild type metabolism is viable. Boxes span the 25-th to 75-th percentile, whiskers indicate maxima and minima, and red '+' signs show outliers.

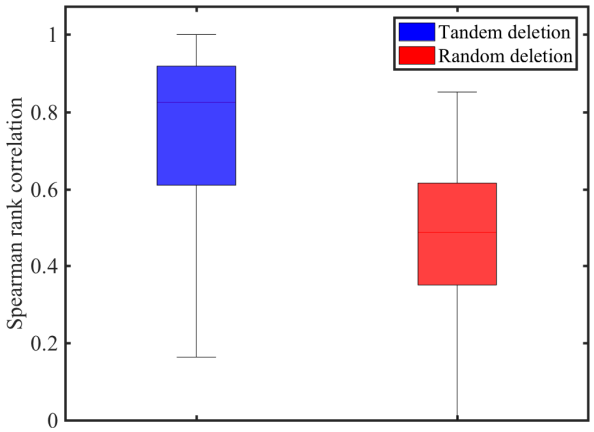



**Figure S4: Higher correlation among carbon sources for tandem robustness as compared to random robustness.** The same metabolism may show different robustness to tandem or random deletions of a given number of genes, depending on the carbon source environment in which this robustness is evaluated. We computed the robustness ($R_{tandem}$) of all our 55 prokaryotic metabolisms to tandem deletions of five genes on all carbon sources on which these metabolisms are viable, and then determined Spearman's rank correlation coefficients between $R_{tandem}$ on these carbon sources, for all pairs of metabolisms. We performed an analogous calculation for robustness ($R_{random}$) to random deletions of five genes. Boxes indicate the distribution of Spearman's rank correlation coefficient of robustness to tandem deletions of five genes ($R_{tandem}$, blue) and random deletions of five genes ($R_{random}$, red) for those carbon sources on which both metabolisms in a given pair are viable. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. In box-whisker plots, boxes span the 25-th to 75-th percentile, whiskers indicate maxima and minima, and red '+' signs show outliers.

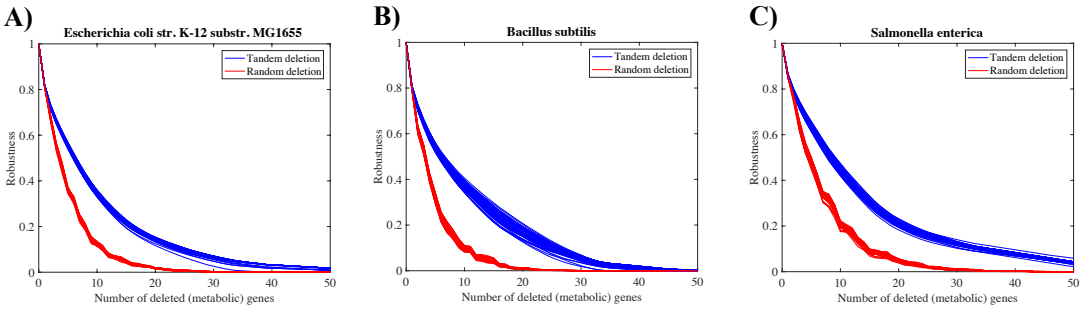



**Figure S5: Higher variability among carbon sources in robustness to tandem deletion than random deletion.** Each panel shows **A)** 97 blue curves and 97 red curves, **B)** 68 blue curves and 68 red curves, and **C)** 87 blue curves and 87 red curves indicating robustness to tandem (blue) and random (red) deletion for each of the A) 97, B) 68, and C) 87 carbon sources on which A) *Escherichia coli* K-12 G1655 *(iJO1366)*, B) *Bacillus subtillis* and C) *Salmonella enterica* are viable as a function of the number of deleted metabolic genes (horizontal axis). Each curve is obtained by a linear interploation betrween 50 data points.

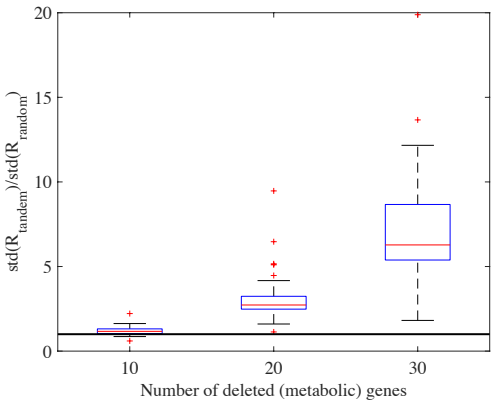



**Figure S6: Excess variability in robustness to tandem deletions.** The vertical axis shows the excess variability in robustness to tandem deletion among different carbon sources defined as the ratio of the standard deviation of robustness to tandem deletion (among different carbon sources) and the standard deviation of robustness to random deletion ($Std(R_{tandem})/Std(R_{random})$), for all 55 bacterial genomes, using three different number of deleted genes (horizontal axis). Boxes span the 25-th to 75-th percentile, whiskers indicate maxima and minima, and red '+' signs show outliers.

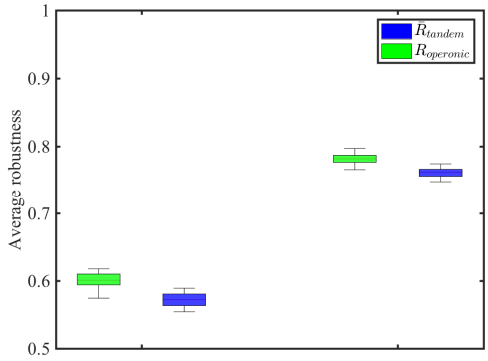



**Figure S7: Robustness to operon deletion.** Boxplots of robustness to operon deletion ($R_{Operonic}$, blue) and of average robustness to tandem deletion of the same length ($\bar{R}_{tandem}$, red, see text S1), computed for 52 bacterial species. For the boxplots of the left-hand side, robustness is defined based on retaining viability on all carbon sources on which the wild-type genome is viable, while for the boxplots of the right hand side robustness is defined based on retaining viability on at least one carbon source.

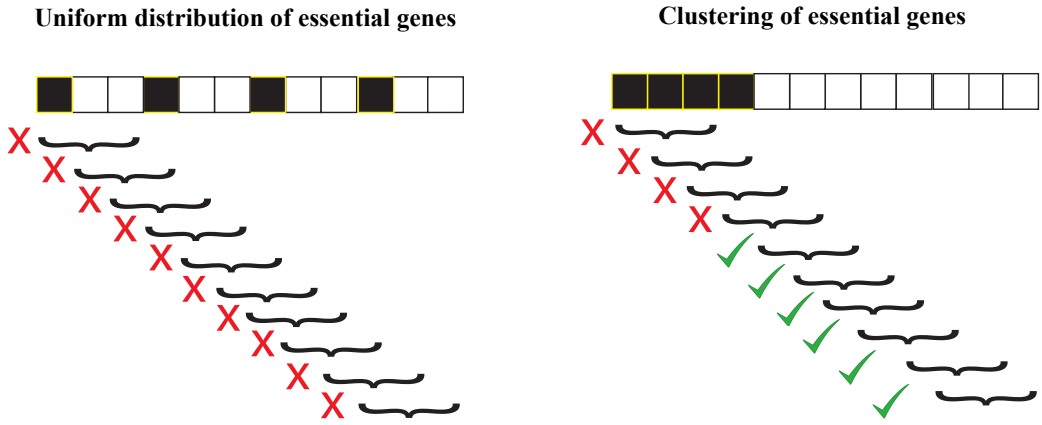



**Figure S8: Clustering of essential genes can enhance robustness to tandem gene deletion**. The figure illustrates two different hypothetical genome organizations, each of which has the same number of genes (12) and the same number of essential genes (4, shown as black boxes). Whereas in the left genome the essential genes are uniformly distributed, in the right genome they are clustered (i.e. conentrated in one region of the genome). Each curly bracket indicates tandem deletions of 3 specific consecutive genes. Red crosses and green check marks indicate whether any one deletion would disrupt or preserve viability. In the left genome, each tandem deletion includes one essential gene and will thus disrupt viability. In contrast, for the right genome, only the left-most four deletional variants include an essential gene, implying that 60% of deletions preserve viability, such that the robustness to deletions is 0.6. Note that we have made the simplifying assumption that simultaneous deletion of multiple non-essential genes is not lethal.

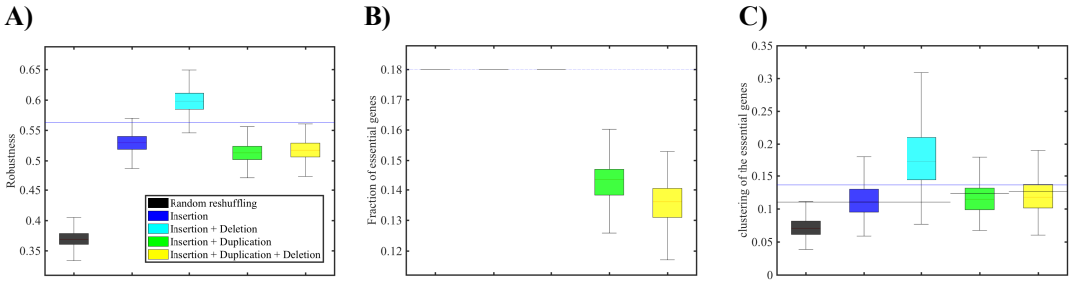



**Figure S9: Emergence of essential gene clusters through gradual genomic expansion from a minimal towards a full-sized genome.** Data is based on 100 genomes "grown" from a minimal genome that is viable on glucose and derived from *Escherichia coli* K-12 G1655 *(iJO1366)*, towards a final genome of equal size as the wild-type *Escherichia coli K-12* genome. We simulated genome "growth" in four different ways (see legend): insertion of genes alone (blue), insertion + deletion (cyan), insertion + duplication (green), and insertion + duplication + deletion (yellow). As a control we generated 100 genomes obtained by random reshuffling of the wild-type *Escherichia coli K-12* genome (black). Vertical axes indicate **A)** robustness to tandem deletions of five genes, **B)** fractions of essential genes, and **C)** clustering of essential genes in the final 100 full-sized genomes, as indicated by Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line in panel A) indicates metabolic robustness of the wild type *Escherichia coli K-12* genome to tandem deletions of five metabolic genes in glucose minimal medium. The blue horizontal line in C) shows the clustering of essential metabolic genes in the wild type *Escherichia coli K-12* genome, as computed by Kuiper's statistics. The black horizontal line in panel C shows the minimal clustering above which the essential genes in a genome are considered significantly clustered (i.e. above which the P-value of the Kuiper's test is below 0.05). Where genome "growth" includes gene duplication, the numbers of essential genes is lowered (panel B), and the minimum clustering threshold increases (panel C).

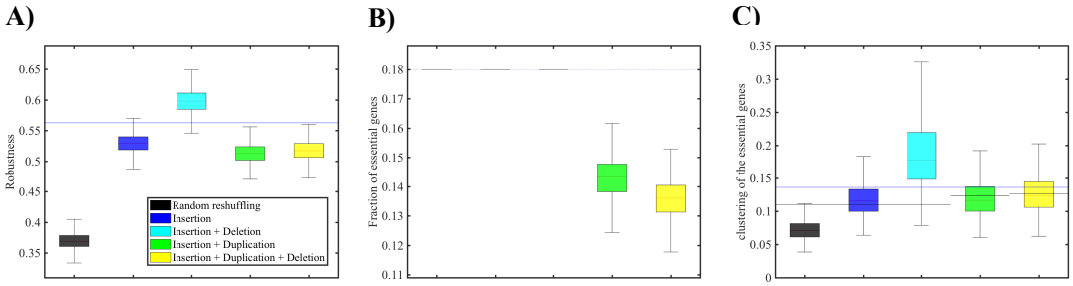



**Figure S10: Emergence of essential gene clusters through gradual genomic expansion from a minimal towards a full-sized genome (using randomly reshuffled minimal genomes).** Data is based on 100 genomes "grown" from a minimal genome that is viable on glucose and derived from *Escherichia coli* K-12 G1655 *(iJO1366)*, towards a final genome of equal size as the wild-type *Escherichia coli K-12* genome. Importantly, in this analysis, we have reshuffled the relative ordering of the genes in the minimal genome. We simulated genome "growth" in four different ways (see legend): insertion of genes alone (blue), insertion + deletion (cyan), insertion + duplication (green), and insertion + duplication + deletion (yellow). As a control we generated 100 genomes obtained by random reshuffling of the wild-type *Escherichia coli K-12* genome (black). Vertical axes indicate **A)** robustness to tandem deletions of five genes, **B)** fractions of essential genes, and **C)** clustering of essential genes in the final 100 full-sized genomes, as indicated by the Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line in panel A) indicates metabolic robustness of the wild type *Escherichia coli K-12* genome to tandem deletions of five metabolic genes in glucose minimal medium. The blue horizontal line in C) shows the clustering of essential metabolic genes in the wild type *Escherichia coli K-12* genome, as computed by Kuiper's statistics. The black horizontal line in panel C shows the minimal clustering above which the essential genes in a genome are considered significantly clustered (i.e. above which the P-value of the Kuiper's test is below 0.05). Where genome "growth" includes gene duplication, the number of essential genes is lowered (panel B), and the minimum clustering threshold increases (panel C).

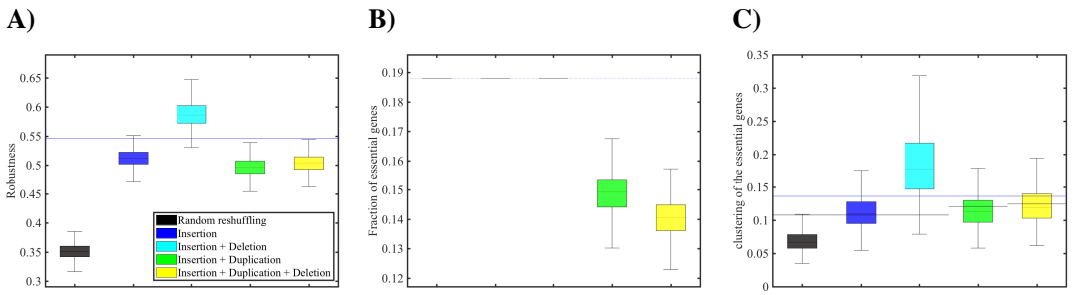



**Figure S11: Emergence of essential gene clusters through gradual genomic expansion from a minimal towards a full-sized genome (using acetate as the carbon source).** Data is based on 100 genomes "grown" from a minimal genome that is viable on acetate and derived from *Escherichia coli* K-12 G1655 *(iJO1366)*, towards a final genome of equal size as the wild-type *Escherichia coli K-12* genome. We simulated genome "growth" in four different ways (see legend): insertion of genes alone (blue), insertion + deletion (cyan), insertion + duplication (green), and insertion + duplication + deletion (yellow). As a control we generated 100 genomes obtained by random reshuffling of the wild-type *Escherichia coli K-12* genome (black). Vertical axes indicate **A)** robustness to tandem deletions of five genes, **B)** fractions of essential genes, and **C)** clustering of essential genes in the final 100 full-sized genomes, as indicated by the Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line in panel A) indicates metabolic robustness of the wild type *Escherichia coli K-12* genome to tandem deletions of five metabolic genes in acetate minimal medium. The blue horizontal line in C) shows the clustering of essential metabolic genes in the wild type *Escherichia coli K-12* genome, as computed by Kuiper's statistics. The black horizontal line in panel C shows the minimal clustering above which the essential genes in a genome are considered significantly clustered (i.e. above which the P-value of the Kuiper's test is below 0.05). Where genome "growth" includes gene duplication, the number of essential genes is lowered (panel B), and the minimum clustering threshold increases (panel C).

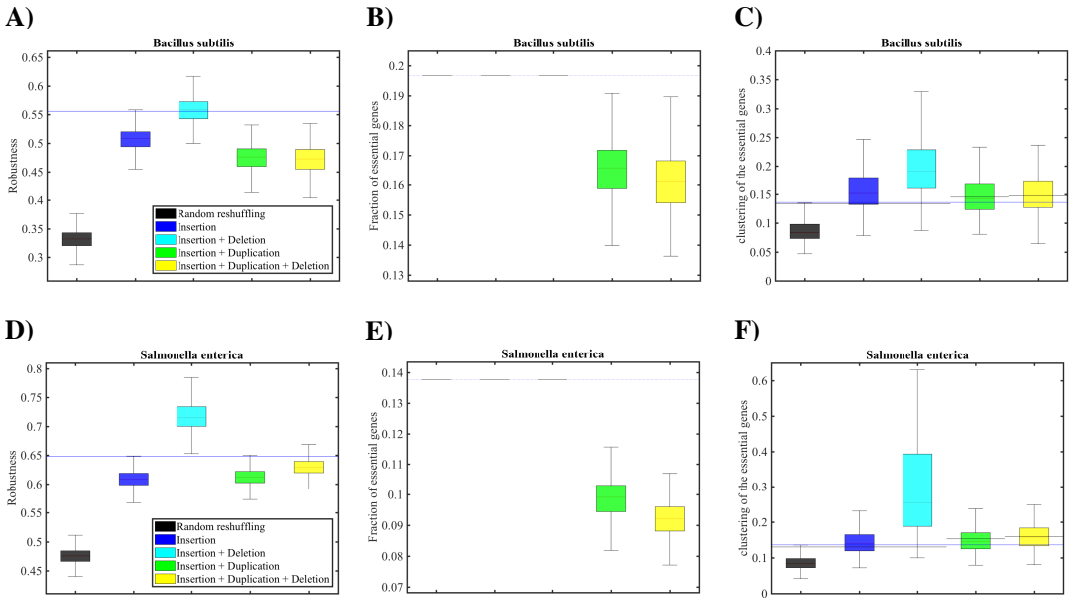



**Figure S12: Emergence of essential gene clusters through gradual genomic expansion from a minimal towards a full-sized genome.** Data in panels A-C is based on 100 genomes "grown" from a minimal genome that is viable on glucose and derived from *Bacillus subtilis*, towards a final genome of equal size as the wild-type *Bacillus subtilis* genome, and the data in panels D-F is based on 100 genomes "grown" from a minimal genome that is viable on glucose and derived from *Salmonella enterica*, towards a final genome of equal size as the wild-type *Salmonella enterica* genome. We simulated genome "growth" in four different ways (see legend): insertion of genes alone (blue), insertion + deletion (cyan), insertion + duplication (green), and insertion + duplication + deletion (yellow). As a control we generated 100 genomes obtained by random reshuffling of the wild-type genome (black). Vertical axes indicate in panels **A)** and **D)** robustness to tandem deletions of five genes, in panels **B)** and **E)** fractions of essential genes, and in panels **C)** and **F)** clustering of essential genes in the final 100 full-sized genomes, as indicated by the Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line in panels A and D indicates metabolic robustness of the wild type *Bacillus subtilis* and *Salmonella enterica* genome to tandem deletions of five metabolic genes in glucose minimal medium. The blue horizontal line in panels C and F shows the clustering of essential metabolic genes in the wild type *Escherichia coli K-12* genome, as computed by Kuiper's statistics. The black horizontal line in panels C and F shows the minimal clustering above which the essential genes in a genome are considered significantly clustered (i.e. above which the P-value of the Kuiper's test is below 0.05). Where genome "growth" includes gene duplication, the number of essential genes is lowered (panels B and E), and the minimum clustering threshold increases (panels C and F).

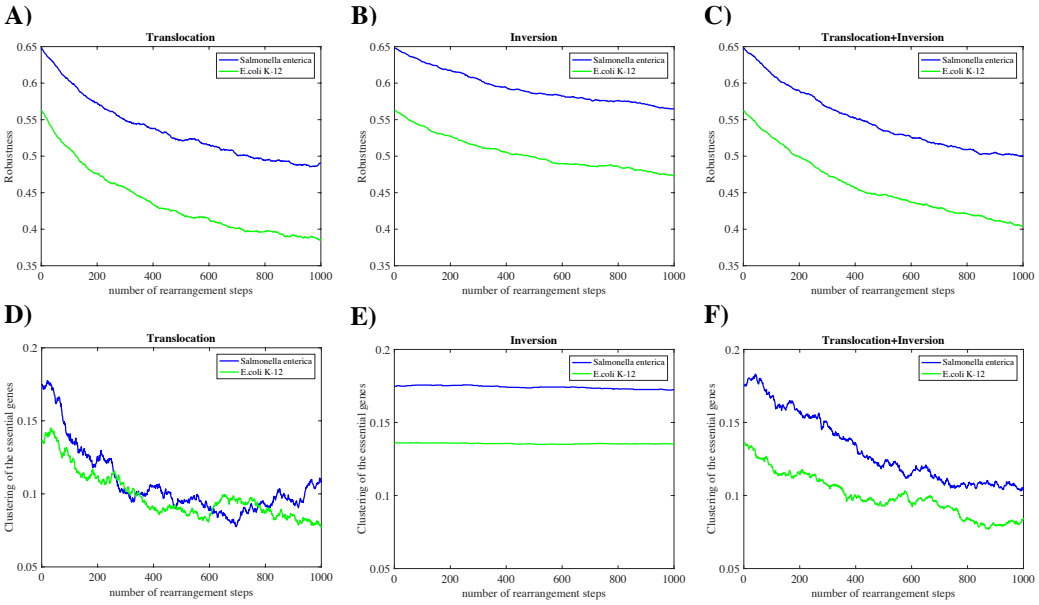



**Figure S13: Genome rearrangement can reduce deletional robustness by disrupting the clusters of essential genes (on glucose).** In each panel, the horizontal axis shows the number of steps in a simulated genome rearrangement process applied independently to 100 initial genomes derived from the wild-type genomes of two organisms (see legend). In each step, each genome is subjected to a genome rearrangement event (translocation (panels A and D), inversion (panels B and E), and translocation or inversion (panels C and F); see methods). The vertical axes in panels A-C show the average robustness to tandem deletions of five genes. In panels D-F they show the average clustering of essential metabolic genes, as computed by Kuiper's statistics, averaged over all 100 genomes. All simulation data reported are based on minimal media containing glucose as the sole carbon source.

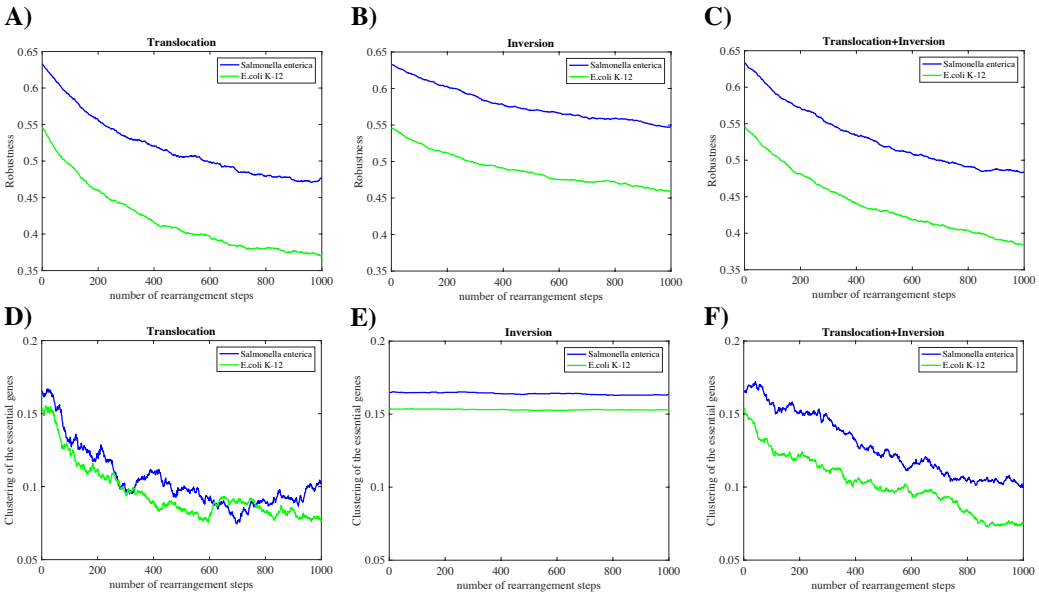



**Figure S14: Genome rearrangement can reduce deletional robustness by disrupting the clusters of essential genes (on acetate).** In each panel, the horizontal axis shows the number of steps in a simulated genome rearrangement process applied independently to 100 initial genomes derived from the wild-type genomes of two organisms (see legend). In each step, each genome is subjected to a genome rearrangement event (translocation (panels A and D), inversion (panels B and E), and translocation or inversion (panels C and F); see methods). The vertical axes in panels A-C show the average robustness to tandem deletions of five genes, and in panels D-F they show the average clustering of essential metabolic genes, as computed by Kuiper's statistics, averaged over all 100 genomes. All simulation data reported are based on minimal media containing acetate as the sole carbon source.

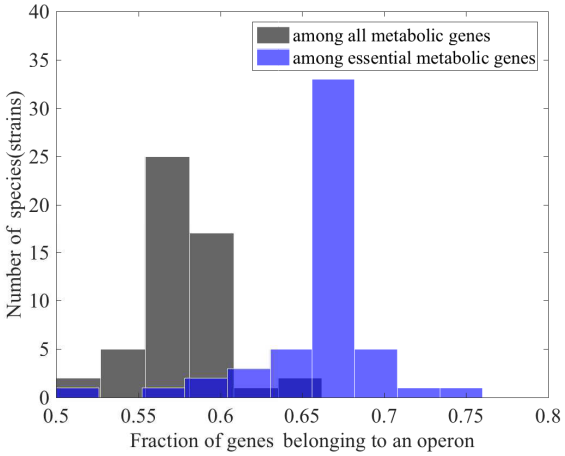



**Figure S15: Operons and conditionally essential genes.** Histogram of the fraction of all metabolic genes (black), and the fraction of conditionally essential metabolic genes (blue) which belong to an operon among the 52 species (strains) used in this analysis. We consider a metabolic gene as conditionally essential, if its deletion results in losing viability on at least one of the carbon sources on which the wild type metabolism is viable.

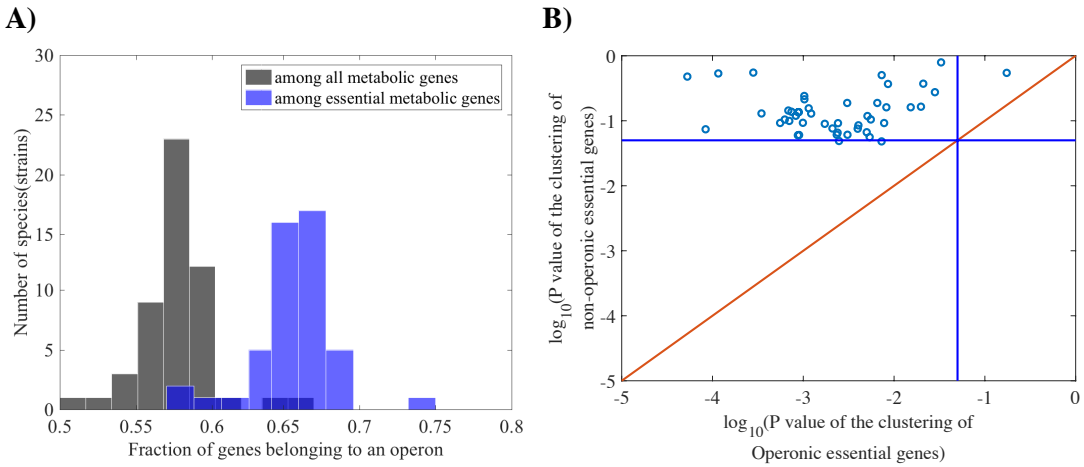



**Figure S16: Operons and the essential genes (on glucose as minimal media). A)** Histogram of the fraction of metabolic genes (black), and the fraction of essential metabolic genes (blue) which belong to an operon among the 52 species (strains) used in this analysis is shown. We consider a metabolic gene as essential, if its deletion results in losing viability on glucose. **B)** In this analysis, we subdivided all strictly essential genes in each of 52 metabolic genomes into two groups i) those belonging to an operon and ii) those not belonging to an operon. For each genome, we determined the extent of gene clustering using the *P*-value generated by Kuiper's test. Each circle in this figure corresponds to a given species or strain. The horizontal and vertical axes show the extent of clustering for essential genes that are part of an operon and not part of an operon, respectively. The blue lines correspond to a significance threshold of $P$=0.05 ($-log_{10}0.05$), and the red line is the identity line. Note that operonic essential genes are significantly clustered in more genomes. Whereas only 2 of the 52 genomes (3.84%) show evidence for clustering of essential genes outside operons (at $P$=0.05), essential genes in operons are clustered in all 52 genomes. Moreover, in all 52 genomes, operonic genes are more significantly clustered (lower *P*-value) than genes outside operons.

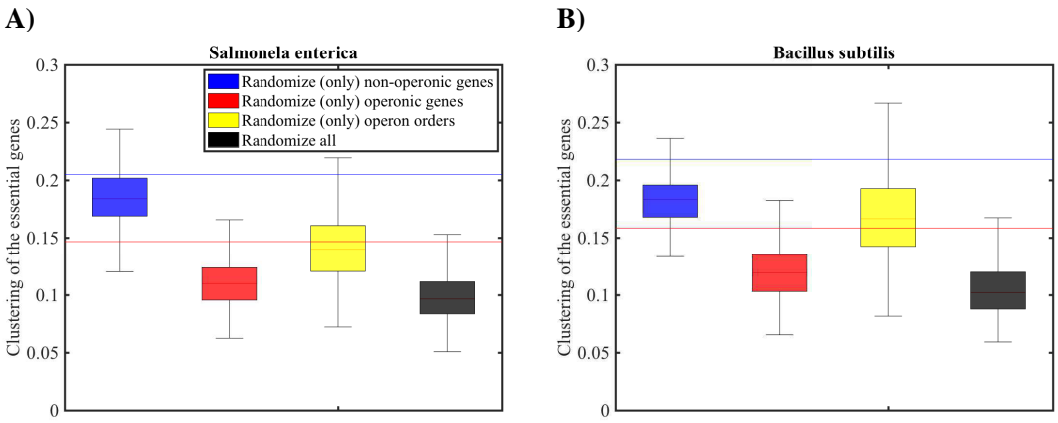



**Figure S17: Impact of operonic genes and operon orders on the clustering of essential genes.** Data in this figure are based on partially or completely randomized **A)** *Salmonella enterica* genome, and **B)** *Bacillus subtilis* genome, with randomized orders of i) all genes (black), ii) non-operonic genes (blue), iii) operonic genes (red), and iv) entire operons (yellow, without changing the intra-operonic gene orders). The vertical axes show the extent of clustering of strictly essential genes in the 100 randomized genomes, as indicated by the Kuiper's test statistic. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. The blue horizontal line indicates the clustering of strictly essential genes in the corresponding wild type genomes as indicated by the Kuiper's test statistic, and the red horizontal line shows a minimal threshold of significant clustering (i.e. where Kuiper's test yields $P<0.05$).

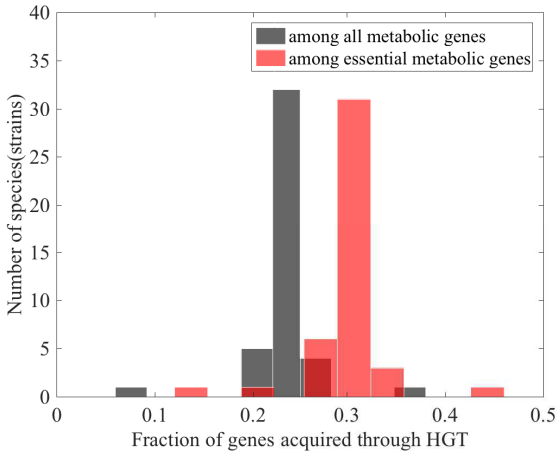



**Figure S18: Horizontal gene transfer and the conditionally essential genes. A)** Histogram of the fraction of metabolic genes (black), and the fraction of conditionally essential metabolic genes (red) which have been acquired through horizontal gene transfer (HGT) among the 43 species (strains) used in this analysis. We consider a metabolic gene as conditionally essential, if its deletion results in losing viability on at least one of the carbon sources on which the wild type metabolism is viable.

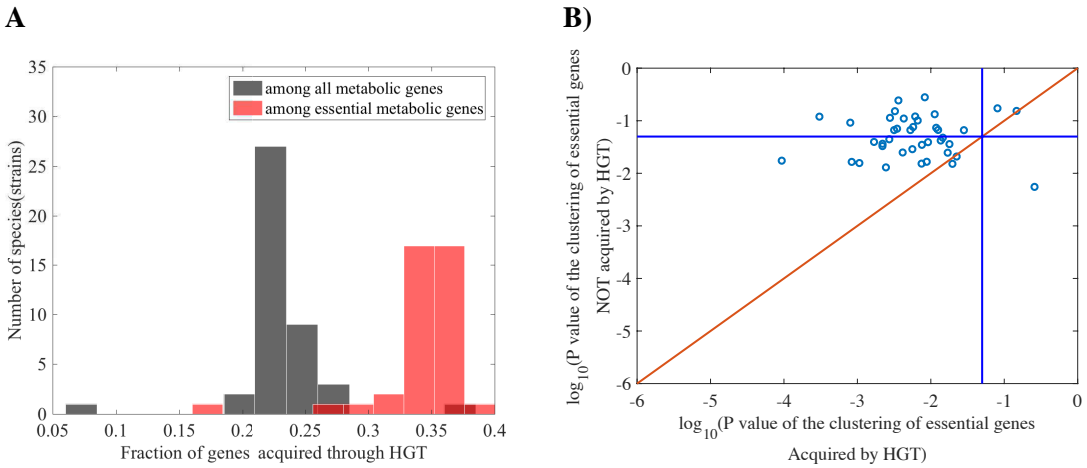



**Figure S19: Horizontal gene transfer and the essential genes (on glucose as minimal media). A)** Histogram of the fraction of metabolic genes (black), and the fraction of essential metabolic genes (red), which have been acquired through horizontal gene transfer (HGT) among the 43 species (strains) used in this analysis. We consider a metabolic gene as essential if its deletion results in losing viability on glucose. **B)** In this analysis, we subdivided all the essential genes in each metabolic genome into two groups: i) those acquired by horizontal gene transfer (HGT) and ii) those not acquired by horizontal gene transfer. For each of the 43 genomes, and separetely for essential genes (on glucose) in each of the two groups, we determined the extent of gene clustering using the P-value generated by Kuiper's test. Each circle in this figure corresponds to a given species or strain. The horizontal and vertical axes show the extent of clustering for essential genes acquired and not acquired by HGT, respectively. The blue lines correspond to a significance threshold of $P$=0.05 ($-log_{10}0.05$), and the red line is the identity line. Note that horizontally transferred essential genes show greater clustering in the vast majority of genomes. Whereas the clustering of horizontally transferred essential genes is significant at $P$=0.05 in 40 among 43 genomes (93.02%), that of not horizontally transferred essential genes is significant only in 21 genomes (48.83%). Moreover, horizontally transferred essential genes are more clustered (lower P-value) than not horizontally transferred essential genes in 40 genomes (93.02%).

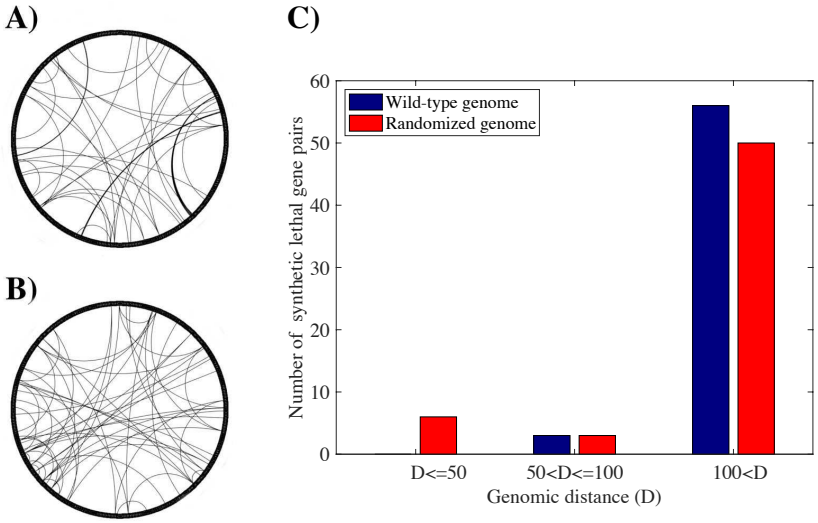



**Figure S20: Repulsion of synthetic lethal genes in the *E. coli 083:H1* genome. A)** Circos plot of *Escherichia coli K-12 083:H1* genome, in which metabolic genes are arranged according to their order in the genome. An arc connects two genes if they form an unconditionally synthetic lethal pair. **B)** Same as A, but for randomized gene order. Note the many short-ranged synthetic lethality interactions after gene order randomization. **C)** Barplot of the genomic distance (in number of intervening genes) between unconditionally synthetic lethal metabolic gene pairs in the wild-type (blue) and randomized (yellow) *Escherichia coli K-12 083:H1* genome. Note the lack of short-distance synthetic lethal pairs with fewer than 5o intervening genes in the wild type genome.

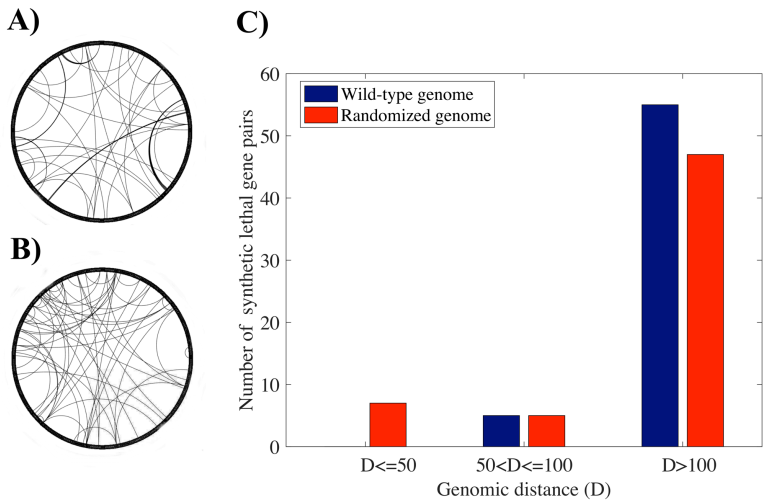



**Figure S21: Repulsion of synthetic lethal genes in the *Shigella flexneri* genome. A)** Circos plot of *Shigella flexneri* genome, in which metabolic genes are arranged according to their order in the genome. An arc connects two genes if they form an unconditionally synthetic lethal pair. **B)** Same as A, but for randomized gene order. Note the many short-ranged synthetic lethality interactions after gene order randomization. **C)** Barplot of the genomic distance (in number of intervening genes) between unconditionally synthetic lethal metabolic gene pairs in the wild-type (blue) and randomized (yellow) *Shigella flexneri* genome. Note the lack of short-distance synthetic lethal pairs with fewer than 5o intervening genes in the wild type genome.

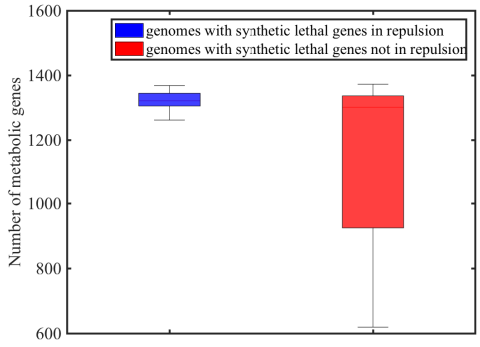



**Figure S22:** Metabolic genomes in which strictly synthetic lethal genes are in repulsion tend to be larger than those in which strictly synthetic lethal genes are not in repulsion (t-test P-value<$10^{-4}$). In none of our genomes with fewer than 1200 metabolic genes were synthetically lethal genes in repulsion.

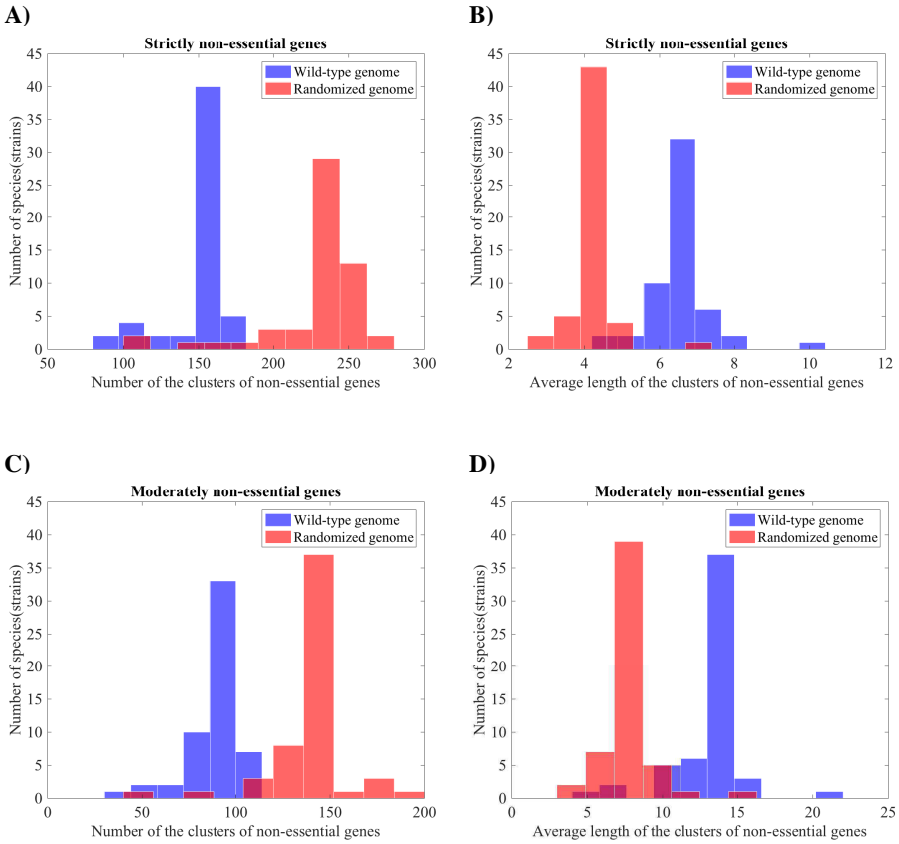



**Figure S23: Long clusters of non-essential genes in bacterial genomes.** Histogram of the number of clusters of **A)** strictly non-essential metabolic genes, and **C)** conditionally non-essential metabolic genes, and the average length of clusters of **B)** strictly non-essential metabolic genes and **D)** conditionally non–essential metabolic genes, among the 55 wild-type bacterial genomes (blue) and the corresponding 55 randomized genomes (red). We consider a metabolic gene as conditionally non-essential if its deletion does not abolish viability on at least one carbon source, and consider a metabolic gene as strictly non-essential if its deletion does not abolish viability on any carbon source. A *cluster of conditionally non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that intervene between two successive but non-adjacent strictly essential metabolic genes. Likewise, a *cluster of strictly non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that intervene between two successive but non-adjacent conditionally essential metabolic genes.

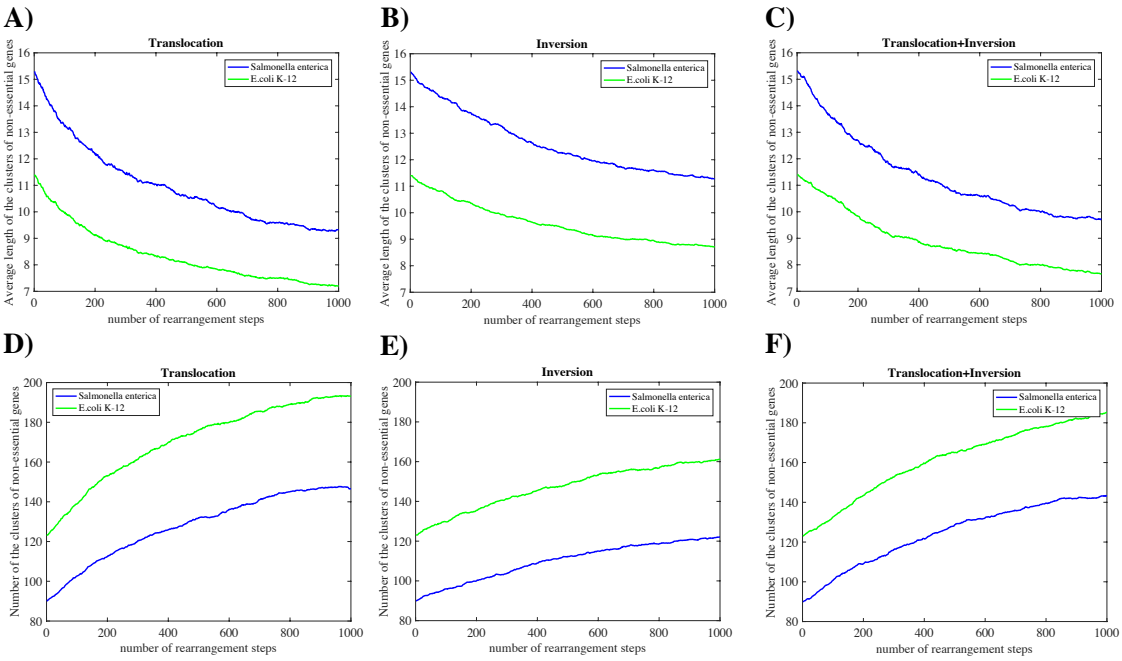



**Figure S24: Genome rearrangement can reduce deletional robustness by shrinking the clusters of non-essential genes (on glucose).** In each panel, the horizontal axis shows the number of steps in a simulated genome rearrangement process applied independently to 100 initial genomes derived from the wild-type genomes of two organisms (see legend). In each step, each genome is subjected to a genome rearrangement event (translocation (panels **A** and **D**), inversion (panels **B** and **E**), and translocation or inversion (panels **C** and **F**); see methods). The vertical axes in panels A-C show the average length of the clusters of non-essential genes. In panels D-F they show the number of the clusters of non-essential genes, averaged over all 100 genomes. All simulation data reported are based on minimal media containing glucose as the sole carbon source. In this analysis, a cluster of non-essential genes are defined as a set of metabolic genes that are not essential on glucose and intervene between two nearest metabolic genes that are essential on glucose.

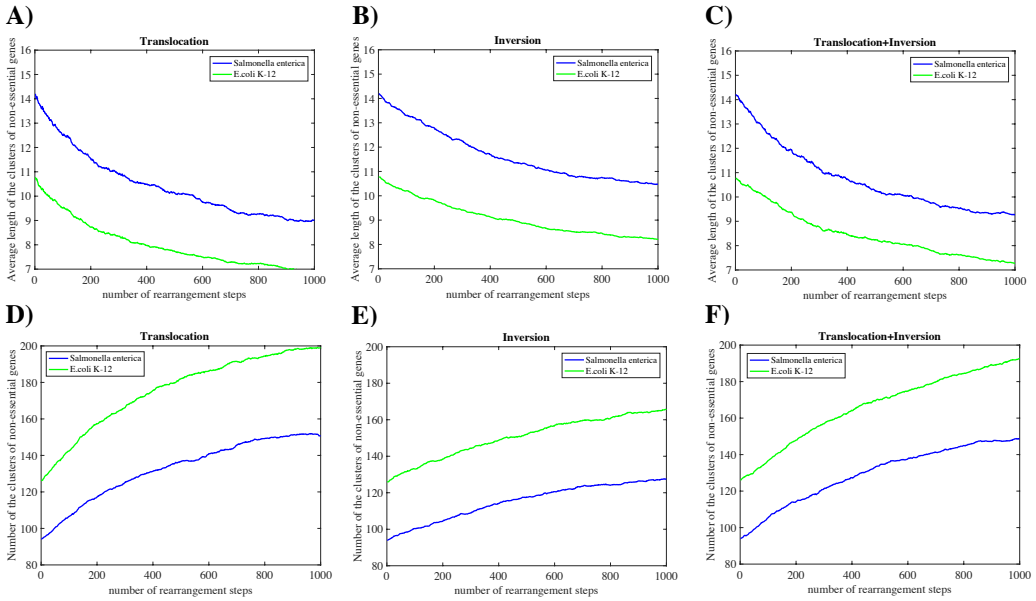



**Figure S25: Genome rearrangement can reduce deletional robustness by shrinking the clusters of non-essential genes (on acetate).** In each panel, the horizontal axis shows the number of steps in a simulated genome rearrangement process applied independently to 100 initial genomes derived from the wild-type genomes of two organisms (see legend). In each step, each genome is subjected to a genome rearrangement event (translocation (panels **A** and **D**), inversion (panels **B** and **E**), and translocation or inversion (panels **C** and **F**); see methods). The vertical axes in panels A-C show the average length of the clusters of non-essential genes. In panels D-F show the number of the clusters of non-essential genes, averaged over all 100 genomes. All simulation data reported are based on minimal media containing acetate as the sole carbon source. In this analysis, a cluster of non-essential genes are defined as a set of metabolic genes that are not essential on acetate and intervene between two nearest metabolic genes that are essential on acetate.

# Supplementary tables:

| Index | Carbon source | Index | Carbon source |
|---|---|---|---|
| 1 | D-Glucose | 52 | L-Tryptophan |
| 2 | Uracil | 53 | Maltose |
| 3 | Acetoacetate | 54 | L-Asparagine |
| 4 | 3-(3-hydroxy-phenyl)propionate | 55 | L-Lactate |
| 5 | N-Acetyl-D-glucosamine | 56 | (S)-Propane-1 |
| 6 | Acetaldehyde | 57 | D-Ribose |
| 7 | N-Acetyl-D-mannosamine | 58 | Sucrose |
| 8 | L-Cysteine | 59 | Thymidine |
| 9 | 2-Dehydro-3-deoxy-D-gluconate | 60 | D-serine |
| 10 | Tetradecanoate (n-C14:0) | 61 | D-Galactose |
| 11 | N-Acetylneuraminate | 62 | Lactose |
| 12 | L-Glutamate | 63 | L-Malate |
| 13 | Uridine | 64 | L-Aspartate |
| 14 | Xanthine | 65 | Putrescine |
| 15 | L-Arginine | 66 | D-Glucose 6-phosphate |
| 16 | L-Alanine | 67 | Phenylpropanoate |
| 17 | Glycolate | 68 | Butyrate (n-C4:0) |
| 18 | Guanine | 69 | Octadecanoate (n-C18:0) |
| 19 | Glycine | 70 | Trehalose |
| 20 | 4-Aminobutanoate | 71 | L-Histidine |
| 21 | L-Glutamine | 72 | Pyruvate |
| 22 | Adenine | 73 | D-Mannitol |
| 23 | Guanosine | 74 | Citrate |
| 24 | Glycerol 3-phosphate | 75 | L-tartrate |
| 25 | D-Glucuronate | 76 | L-Threonine |
| 26 | Glycerol | 77 | Ornithine |
| 27 | Hexadecanoate (n-C16:0) | 78 | Maltopentaose |
| 28 | Adenosine | 79 | Maltotriose |
| 29 | D-Glyceraldehyde | 80 | Maltohexaose |
| 30 | D-Glucosamine | 81 | L-Rhamnose |
| 31 | D-Galacturonate | 82 | Succinate |
| 32 | D-Galactonate | 83 | D-Mannose |
| 33 | D-Glucarate | 84 | Cytidine |
| 34 | Hypoxanthine | 85 | D-Sorbitol |
| 35 | D-Gluconate | 86 | Deoxyadenosine |
| 36 | 2-Oxoglutarate | 87 | Maltotetraose |
| 37 | Galactitol | 88 | Melibiose |
| 38 | 3-hydroxycinnamic acid | 89 | D-Mannose 6-phosphate |
| 39 | Allantoin | 90 | Deoxycytidine |
| 40 | D-Galactarate | 91 | L-Fucose |
| 41 | D-Xylose | 92 | L-Serine |
| 42 | L-Idonate | 93 | D-Fructose |
| 43 | Acetate | 94 | Deoxyguanosine |
| 44 | Xanthosine | 95 | Dihydroxyacetone |
| 45 | AMP | 96 | Fumarate |
| 46 | L-Isoleucine | 97 | Cytosine |
| 47 | Inosine | 98 | Deoxyinosine |
| 48 | L-Arabinose | 99 | D-Alanine |
| 49 | L-Valine | 100 | Deoxyuridine |
| 50 | D-Lactate | 101 | Formate |
| 51 | L-Proline | 102 | Ethanol |

**Table S1: List of carbon sources used in this study.**

| Species (strains) | Number of operons | Average number of genes per operon | Average robustness to operon deletion (strict definition) | Average (normalized) robustness of genome (strict definition) | Excess robustness to operon deletion (strict definition) | Average robustness to operon deletion (moderate definition) | Average (normalized) robustness of genome (moderate definition) | Excess robustness to operon deletion (moderate definition) |
|---|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 241 | 3.07 | 0.6 | 0.56 | 1.08 | 0.8 | 0.77 | 1.03 |
| Methanosarcina barkeri str. Fusaro | 91 | 2.75 | 0.77 | 0.62 | 1.23 | 0.77 | 0.62 | 1.23 |
| Geobacter metallireducens GS-15 | 203 | 3.12 | 0.66 | 0.56 | 1.17 | 0.66 | 0.56 | 1.17 |
| *E. coli* APEC O1 | 247 | 2.94 | 0.59 | 0.57 | 1.04 | 0.79 | 0.76 | 1.04 |
| *E. coli* BL21(DE3) [iB21 1397] | 277 | 3.19 | 0.6 | 0.57 | 1.07 | 0.78 | 0.75 | 1.03 |
| *E. coli* BW2952 | 247 | 3.14 | 0.54 | 0.52 | 1.03 | 0.72 | 0.69 | 1.03 |
| *E. coli* CFT073 | 260 | 2.8 | 0.77 | 0.72 | 1.06 | 0.77 | 0.72 | 1.06 |
| *E. coli* O127:H6 | 244 | 3.01 | 0.6 | 0.57 | 1.05 | 0.78 | 0.76 | 1.03 |
| *E. coli* 042 | 252 | 3.07 | 0.6 | 0.56 | 1.06 | 0.78 | 0.76 | 1.03 |
| *E. coli* 55989 | 257 | 3.05 | 0.61 | 0.57 | 1.06 | 0.78 | 0.76 | 1.02 |
| *E. coli* ABU 83972 | 247 | 3.07 | 0.6 | 0.57 | 1.04 | 0.79 | 0.76 | 1.04 |
| *E. coli* B str. REL606 | 256 | 3 | 0.61 | 0.57 | 1.06 | 0.78 | 0.77 | 1.02 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 256 | 3.05 | 0.61 | 0.58 | 1.06 | 0.78 | 0.76 | 1.02 |
| *E. coli* BL21(DE3) [iECD1391] | 253 | 3.02 | 0.6 | 0.57 | 1.06 | 0.77 | 0.76 | 1.02 |
| *E. coli* DH1 [iEcDH1 1363] | 260 | 3.09 | 0.61 | 0.58 | 1.04 | 0.78 | 0.77 | 1.01 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 255 | 3.1 | 0.6 | 0.61 | 0.98 | 0.78 | 0.79 | 0.99 |
| *E. coli* E24377A | 260 | 2.98 | 0.61 | 0.58 | 1.05 | 0.79 | 0.77 | 1.02 |
| *E. coli* ED1a | 242 | 3 | 0.57 | 0.56 | 1.02 | 0.78 | 0.76 | 1.03 |
| *E. coli* O157:H7 | 240 | 2.95 | 0.59 | 0.56 | 1.05 | 0.77 | 0.76 | 1.02 |
| *E. coli* HS | 253 | 3.06 | 0.6 | 0.57 | 1.06 | 0.78 | 0.76 | 1.02 |
| *E. coli* NA114 | 245 | 3.03 | 0.6 | 0.57 | 1.05 | 0.78 | 0.76 | 1.03 |
| *E. coli* O103:H2 str. 12009 | 250 | 3.06 | 0.6 | 0.58 | 1.04 | 0.78 | 0.76 | 1.02 |
| *E. coli* O111:H- str. 11128 | 251 | 3.05 | 0.61 | 0.58 | 1.05 | 0.78 | 0.76 | 1.02 |
| *E. coli* O26:H11 str. 11368 | 261 | 3.04 | 0.61 | 0.58 | 1.06 | 0.79 | 0.77 | 1.02 |
| *E. coli* IHE3034 | 242 | 3.01 | 0.59 | 0.57 | 1.04 | 0.79 | 0.76 | 1.03 |
| *E. coli* ATCC 8739 | 258 | 3.1 | 0.62 | 0.58 | 1.06 | 0.79 | 0.77 | 1.02 |
| *E. coli* 536 | 248 | 2.97 | 0.59 | 0.57 | 1.04 | 0.79 | 0.76 | 1.03 |
| *E. coli* O157:H7 str. Sakai | 248 | 2.99 | 0.6 | 0.57 | 1.05 | 0.78 | 0.76 | 1.02 |
| *E. coli* S88 | 248 | 3.01 | 0.6 | 0.57 | 1.04 | 0.79 | 0.76 | 1.04 |
| *E. coli* SE11 | 259 | 3.08 | 0.61 | 0.58 | 1.05 | 0.78 | 0.77 | 1.02 |
| *E. coli* SE15 | 251 | 3.07 | 0.53 | 0.51 | 1.04 | 0.73 | 0.7 | 1.05 |
| *E. coli* SMS-3-5 | 260 | 3.08 | 0.6 | 0.57 | 1.05 | 0.78 | 0.76 | 1.03 |
| *E. coli* O157:H7 str. TW14359 | 245 | 3 | 0.6 | 0.57 | 1.04 | 0.77 | 0.76 | 1.01 |
| *E. coli* UMN026 | 259 | 3.05 | 0.61 | 0.57 | 1.08 | 0.79 | 0.76 | 1.03 |
| *E. coli* W [iECW_1372] | 259 | 3.03 | 0.61 | 0.59 | 1.04 | 0.78 | 0.77 | 1.01 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *E. coli* KO11FL | 258 | 3.09 | 0.62 | 0.58 | 1.06 | 0.79 | 0.77 | 1.03 |
| *E. coli* ETEC H10407 | 257 | 3.01 | 0.61 | 0.58 | 1.04 | 0.78 | 0.77 | 1.02 |
| *E. coli* O55:H7 str. CB9615 | 247 | 3.02 | 0.59 | 0.55 | 1.07 | 0.78 | 0.76 | 1.03 |
| *E. coli* K-12 MG1655 [iJO1366] | 262 | 3.12 | 0.56 | 0.52 | 1.07 | 0.74 | 0.71 | 1.04 |
| *E. coli* K-12 MG1655 [iJR904] | 181 | 2.87 | 0.55 | 0.43 | 1.28 | 0.81 | 0.72 | 1.11 |
| *E. coli* LF82 | 246 | 3.02 | 0.6 | 0.58 | 1.03 | 0.78 | 0.77 | 1.01 |
| Mycobacterium tuberculosis H37Rv | 122 | 2.82 | 0.55 | 0.48 | 1.15 | 0.55 | 0.48 | 1.15 |
| *E. coli* O83:H1 str. NRG 857C | 248 | 3.03 | 0.6 | 0.58 | 1.03 | 0.78 | 0.76 | 1.03 |
| Shigella flexneri 2a str. 2457T | 218 | 2.93 | 0.58 | 0.55 | 1.04 | 0.73 | 0.73 | 1 |
| Shigella flexneri 5 str. 8401 | 218 | 2.94 | 0.6 | 0.53 | 1.13 | 0.75 | 0.81 | 0.93 |
| *E. coli* UM146 | 249 | 3 | 0.6 | 0.55 | 1.08 | 0.79 | 0.73 | 1.08 |
| *E. coli* UMNK88 | 259 | 3.03 | 0.62 | 0.57 | 1.07 | 0.78 | 0.76 | 1.03 |
| *E. coli* UTI89 | 259 | 2.95 | 0.61 | 0.57 | 1.06 | 0.79 | 0.75 | 1.06 |
| Klebsiella pneumoniae MGH78578 | 227 | 3 | 0.68 | 0.57 | 1.19 | 0.89 | 0.76 | 1.16 |
| Bacillus subtilis str. 168 | 161 | 3.24 | 0.51 | 0.58 | 0.88 | 0.79 | 0.76 | 1.03 |
| *E. coli* O157:H7 str. EDL933 | 253 | 3.02 | 0.61 | 0.58 | 1.04 | 0.78 | 0.77 | 1.01 |
| Salmonella Typhimurium str. LT2 | 247 | 2.95 | 0.66 | 0.63 | 1.05 | 0.82 | 0.85 | 0.96 |

**Table S2: Operons and deletional robustness.** Each row corresponds to a given species or strain. Columns, from left to right, show species or strain names, the total number of operons in the genome, the average number of genes per operon, the average robustness to operon deletion, average normalized robustness to tandem gene deletion (see text S1), excess robustness to operon deletion, that is, the robustness to operon deletion divided by the average robustness to tandem deletion. In columns 4-6, robustness requires retaining viability on *all* carbon sources. Columns 7 to 9 shows the same information as columns 4-6 respectively but for a less stringent definition of robustness that requires retaining viability on at least one carbon source.

| Species (strain) | Number of strictly essential genes | Number of conditionally essential genes |
|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 144 | 322 |
| Methanosarcina barkeri str. Fusaro | 139 | 139 |
| Geobacter metallireducens GS-15 | 268 | 268 |
| *E. coli* APEC O1 | 160 | 320 |
| *E. coli* BL21(DE3) [iB21 1397] | 164 | 326 |
| *E. coli* BW2952 | 210 | 367 |
| *E. coli* CFT073 | 204 | 204 |
| *E. coli* O127:H6 | 160 | 313 |
| *E. coli* 042 | 159 | 329 |
| *E. coli* 55989 | 159 | 326 |
| *E. coli* ABU 83972 | 160 | 320 |
| *E. coli* B str. REL606 | 159 | 327 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 164 | 326 |
| *E. coli* BL21(DE3) [iECD1391] | 164 | 327 |
| *E. coli* DH1 [iEcDH1 1363] | 159 | 323 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 158 | 322 |
| *E. coli* E24377A | 157 | 324 |
| *E. coli* ED1a | 161 | 319 |
| *E. coli* O157:H7 | 159 | 357 |
| *E. coli* HS | 159 | 336 |
| *E. coli* NA114 | 158 | 320 |
| *E. coli* O103:H2 str. 12009 | 159 | 322 |
| *E. coli* O111:H- str. 11128 | 159 | 346 |
| *E. coli* O26:H11 str. 11368 | 159 | 337 |
| *E. coli* IHE3034 | 160 | 320 |
| *E. coli* ATCC 8739 | 159 | 323 |
| *E. coli* 536 | 160 | 320 |
| *E. coli* O157:H7 str. Sakai | 159 | 323 |
| *E. coli* S88 | 160 | 319 |
| *E. coli* SE11 | 159 | 323 |
| *E. coli* SE15 | 210 | 375 |
| *E. coli* SMS-3-5 | 159 | 328 |
| *E. coli* O157:H7 str. TW14359 | 159 | 322 |
| *E. coli* UMN026 | 158 | 332 |
| *E. coli* W [iECW_1372] | 158 | 326 |
| *E. coli* KO11FL | 158 | 321 |
| *E. coli* ETEC H10407 | 159 | 333 |
| *E. coli* O55:H7 str. CB9615 | 159 | 337 |
| *E. coli* K-12 MG1655 [iJO1366] | 207 | 383 |
| *E. coli* K-12 MG1655 [iJR904] | 134 | 316 |
| *E. coli* LF82 | 182 | 403 |
| Mycobacterium tuberculosis H37Rv | 197 | 197 |
| *E. coli* O83:H1 str. NRG 857C | 160 | 317 |
| Shigella flexneri 2a str. 2457T | 167 | 334 |
| Staphylococcus aureus N315 | 61 | 205 |
| Shigella flexneri 5 str. 8401 | 164 | 320 |
| *E. coli* UM146 | 159 | 319 |
| *E. coli* UMNK88 | 159 | 319 |
| *E. coli* UTI89 | 160 | 320 |
| *E. coli* W [iWFL 1372] | 158 | 326 |
| *E. coli* str. K-12 W3110 | 159 | 329 |
| Klebsiella pneumoniae MGH78578 | 89 | 402 |
| Bacillus subtilis str. 168 | 120 | 288 |
| *E. coli* O157:H7 str. EDL933 | 159 | 322 |
| Salmonella Typhimurium str. LT2 | 140 | 293 |

**Table S3:** Each row corresponds to a bacterial species or strain. Columns, from left to right, show species (strain) name, number of strictly essential metabolic genes, and number of conditionally essential metabolic genes. We consider a metabolic gene strictly essential, if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable, and we consider a metabolic gene conditionally essential, if its deletion abolishes viability on at least one carbon source.

| Species (strain) | Hypothesis rejected | P-value | Kuiper test statistic | Critical value |
|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 1 | 5.48E-03 | 0.1713 | 0.1435 |
| Methanosarcina barkeri str. Fusaro | 0 | 2.53E-01 | 0.1211 | 0.1492 |
| Geobacter metallireducens GS-15 | 1 | 1.84E-05 | 0.1688 | 0.1078 |
| *E. coli* APEC O1 | 1 | 4.16E-05 | 0.2082 | 0.1371 |
| *E. coli* BL21(DE3) [iB21 1397] | 1 | 3.76E-03 | 0.1656 | 0.1354 |
| *E. coli* BW2952 | 1 | 2.31E-04 | 0.1694 | 0.1197 |
| *E. coli* CFT073 | 1 | 5.52E-04 | 0.1687 | 0.1242 |
| *E. coli* O127:H6 | 1 | 2.15E-03 | 0.1733 | 0.1371 |
| *E. coli* 042 | 1 | 7.55E-04 | 0.1839 | 0.1375 |
| *E. coli* 55989 | 1 | 6.16E-04 | 0.1858 | 0.1375 |
| *E. coli* ABU 83972 | 1 | 1.17E-03 | 0.1792 | 0.1371 |
| *E. coli* B str. REL606 | 1 | 3.72E-03 | 0.1682 | 0.1375 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 1 | 6.43E-03 | 0.1599 | 0.1354 |
| *E. coli* BL21(DE3) [iECD1391] | 1 | 2.63E-03 | 0.1692 | 0.1354 |
| *E. coli* DH1 [iEcDH1 1363] | 1 | 6.30E-04 | 0.1856 | 0.1375 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 1 | 2.42E-02 | 0.1474 | 0.138 |
| *E. coli* E24377A | 1 | 6.38E-04 | 0.1855 | 0.1375 |
| *E. coli* ED1a | 1 | 5.07E-03 | 0.164 | 0.1367 |
| *E. coli* O157:H7 | 1 | 9.67E-04 | 0.1816 | 0.1375 |
| *E. coli* HS | 1 | 1.74E-03 | 0.1759 | 0.1375 |
| *E. coli* NA114 | 1 | 2.84E-03 | 0.1716 | 0.138 |
| *E. coli* O103:H2 str. 12009 | 1 | 1.37E-03 | 0.1782 | 0.1375 |
| *E. coli* O111:H- str. 11128 | 1 | 8.53E-04 | 0.1828 | 0.1375 |
| *E. coli* O26:H11 str. 11368 | 1 | 8.09E-04 | 0.1833 | 0.1375 |
| *E. coli* IHE3034 | 1 | 2.31E-03 | 0.1726 | 0.1371 |
| *E. coli* ATCC 8739 | 1 | 1.15E-03 | 0.18 | 0.1375 |
| *E. coli* 536 | 1 | 1.31E-03 | 0.1782 | 0.1371 |
| *E. coli* O157:H7 str. Sakai | 1 | 1.00E-03 | 0.1812 | 0.1375 |
| *E. coli* S88 | 1 | 2.07E-03 | 0.1737 | 0.1371 |
| *E. coli* SE11 | 1 | 6.58E-04 | 0.1852 | 0.1375 |
| *E. coli* SE15 | 1 | 1.51E-03 | 0.1543 | 0.1197 |
| *E. coli* SMS-3-5 | 1 | 7.94E-04 | 0.1835 | 0.1375 |
| *E. coli* O157:H7 str. TW14359 | 1 | 1.15E-03 | 0.1799 | 0.1375 |
| *E. coli* UMN026 | 1 | 1.27E-03 | 0.1785 | 0.1371 |
| *E. coli* W [iECW_1372] | 1 | 2.09E-04 | 0.196 | 0.138 |
| *E. coli* KO11FL | 1 | 5.37E-04 | 0.1876 | 0.138 |
| *E. coli* ETEC H10407 | 1 | 7.23E-04 | 0.1843 | 0.1375 |
| *E. coli* O55:H7 str. CB9615 | 1 | 3.12E-04 | 0.192 | 0.1375 |
| *E. coli* K-12 MG1655 [iJO1366] | 1 | 8.37E-04 | 0.1596 | 0.12 |
| *E. coli* K-12 MG1655 [iJR904] | 1 | 3.15E-02 | 0.1565 | 0.1498 |
| *E. coli* LF82 | 1 | 2.06E-09 | 0.275 | 0.1371 |
| Mycobacterium tuberculosis H37Rv | 0 | 2.95E-01 | 0.097 | 0.123 |
| *E. coli* O83:H1 str. NRG 857C | 1 | 1.59E-03 | 0.1763 | 0.1371 |
| Shigella flexneri 2a str. 2457T | 1 | 1.13E-02 | 0.1533 | 0.135 |
| Staphylococcus aureus N315 | 0 | 6.50E-02 | 0.2198 | 0.2255 |
| Shigella flexneri 5 str. 8401 | 1 | 9.77E-03 | 0.1554 | 0.1354 |
| *E. coli* UM146 | 1 | 7.51E-04 | 0.184 | 0.1375 |
| *E. coli* UMNK88 | 0 | 1.73E-01 | 0.1185 | 0.1375 |
| *E. coli* UTI89 | 1 | 1.30E-03 | 0.1782 | 0.1371 |
| *E. coli* W [iWFL 1372] | 1 | 2.09E-04 | 0.196 | 0.138 |
| *E. coli* str. K-12 W3110 | 1 | 9.73E-04 | 0.1815 | 0.1375 |
| Klebsiella pneumoniae MGH78578 | 1 | 4.24E-02 | 0.1922 | 0.1887 |
| Bacillus subtilis str. 168 | 1 | 4.12E-04 | 0.218 | 0.1583 |
| *E. coli* O157:H7 str. EDL933 | 1 | 3.00E-04 | 0.1923 | 0.1375 |
| Salmonella Typhimurium str. LT2 | 1 | 3.02E-04 | 0.2048 | 0.1465 |

**Table S4: Clustering of the strictly essential genes.** Each row corresponds to a bacterial species or strain. Columns, from left to right, show species (strain) name, whether the null hypothesis of a uniform distribution of strictly essential genes is rejected by Kuiper's test (1) or not (0), the P-value of Kuiper's test, Kuiper's test statistics, and the critical value of this statistic above which the null hypothesis is rejected. In 51 among the 55 species the null hypothesis is rejected, i.e., essential genes are significantly clustered. We consider a metabolic gene strictly essential if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Species (strain) | Hypothesis rejected | P-value | Kuiper test statistic | Critical value |
|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 1 | 2.47E-06 | 0.1615 | 0.0965 |
| Methanosarcina barkeri str. Fusaro | 0 | 2.97E-01 | 0.1151 | 0.146 |
| Geobacter metallireducens GS-15 | 1 | 2.16E-05 | 0.1644 | 0.1056 |
| *E. coli* APEC O1 | 1 | 5.29E-05 | 0.1456 | 0.0968 |
| *E. coli* BL21(DE3) [iB21 1397] | 1 | 1.38E-05 | 0.1516 | 0.0959 |
| *E. coli* BW2952 | 1 | 1.03E-05 | 0.1444 | 0.0904 |
| *E. coli* CFT073 | 1 | 1.50E-04 | 0.1743 | 0.1209 |
| *E. coli* O127:H6 | 1 | 1.73E-04 | 0.1402 | 0.0978 |
| *E. coli* 042 | 1 | 3.57E-06 | 0.1579 | 0.0954 |
| *E. coli* 55989 | 1 | 8.51E-08 | 0.1764 | 0.0959 |
| *E. coli* ABU 83972 | 1 | 8.84E-06 | 0.1554 | 0.0968 |
| *E. coli* B str. REL606 | 1 | 6.29E-06 | 0.1555 | 0.0957 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 1 | 3.51E-05 | 0.1465 | 0.0959 |
| *E. coli* BL21(DE3) [iECD1391] | 1 | 5.90E-06 | 0.1558 | 0.0957 |
| *E. coli* DH1 [iEcDH1 1363] | 1 | 1.78E-06 | 0.1629 | 0.0963 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 1 | 1.33E-05 | 0.1527 | 0.0965 |
| *E. coli* E24377A | 1 | 2.53E-07 | 0.172 | 0.0962 |
| *E. coli* ED1a | 1 | 2.67E-04 | 0.1362 | 0.0969 |
| *E. coli* O157:H7 | 1 | 8.94E-05 | 0.1351 | 0.0917 |
| *E. coli* HS | 1 | 1.56E-06 | 0.1603 | 0.0945 |
| *E. coli* NA114 | 1 | 3.42E-05 | 0.1481 | 0.0968 |
| *E. coli* O103:H2 str. 12009 | 1 | 1.82E-06 | 0.163 | 0.0965 |
| *E. coli* O111:H- str. 11128 | 1 | 2.57E-02 | 0.099 | 0.0931 |
| *E. coli* O26:H11 str. 11368 | 1 | 1.48E-08 | 0.1811 | 0.0943 |
| *E. coli* IHE3034 | 1 | 1.52E-05 | 0.1525 | 0.0968 |
| *E. coli* ATCC 8739 | 1 | 2.33E-06 | 0.1615 | 0.0963 |
| *E. coli* 536 | 1 | 5.57E-06 | 0.1578 | 0.0968 |
| *E. coli* O157:H7 str. Sakai | 1 | 1.55E-06 | 0.1635 | 0.0963 |
| *E. coli* S88 | 1 | 3.53E-05 | 0.1481 | 0.0969 |
| *E. coli* SE11 | 1 | 6.52E-07 | 0.1678 | 0.0963 |
| *E. coli* SE15 | 1 | 5.42E-05 | 0.1345 | 0.0895 |
| *E. coli* SMS-3-5 | 1 | 6.30E-06 | 0.1553 | 0.0956 |
| *E. coli* O157:H7 str. TW14359 | 1 | 2.66E-06 | 0.1611 | 0.0965 |
| *E. coli* UMN026 | 1 | 3.80E-06 | 0.1569 | 0.095 |
| *E. coli* W [iECW_1372] | 1 | 3.52E-07 | 0.1699 | 0.0959 |
| *E. coli* KO11FL | 1 | 1.41E-06 | 0.1645 | 0.0966 |
| *E. coli* ETEC H10407 | 1 | 4.25E-08 | 0.1776 | 0.0949 |
| *E. coli* O55:H7 str. CB9615 | 1 | 6.88E-07 | 0.164 | 0.0943 |
| *E. coli* K-12 MG1655 [iJO1366] | 1 | 2.98E-06 | 0.1473 | 0.0885 |
| *E. coli* K-12 MG1655 [iJR904] | 1 | 2.93E-03 | 0.1209 | 0.0974 |
| *E. coli* LF82 | 1 | 5.38E-04 | 0.1174 | 0.0863 |
| Mycobacterium tuberculosis H37Rv | 0 | 2.95E-01 | 0.097 | 0.123 |
| *E. coli* O83:H1 str. NRG 857C | 1 | 7.04E-06 | 0.1573 | 0.0972 |
| Shigella flexneri 2a str. 2457T | 1 | 6.89E-06 | 0.1534 | 0.0947 |
| Staphylococcus aureus N315 | 1 | 6.73E-04 | 0.1622 | 0.1206 |
| Shigella flexneri 5 str. 8401 | 1 | 1.82E-05 | 0.1515 | 0.0968 |
| *E. coli* UM146 | 1 | 1.71E-05 | 0.1521 | 0.0969 |
| *E. coli* UMNK88 | 1 | 1.43E-05 | 0.1531 | 0.0969 |
| *E. coli* UTI89 | 1 | 6.20E-06 | 0.1573 | 0.0968 |
| *E. coli* W [iWFL 1372] | 1 | 3.52E-07 | 0.1699 | 0.0959 |
| *E. coli* str. K-12 W3110 | 1 | 7.75E-08 | 0.176 | 0.0954 |
| Klebsiella pneumoniae_MGH78578 | 1 | 1.27E-17 | 0.2322 | 0.0864 |
| Bacillus subtilis str. 168 | 0 | 4.62E-01 | 0.0729 | 0.1019 |
| *E. coli* O157:H7 str. EDL933 | 1 | 2.36E-07 | 0.1728 | 0.0965 |
| Salmonella Typhimurium str. LT2 | 1 | 4.15E-04 | 0.1392 | 0.1011 |

**Table S5: Clustering of the conditionally essential genes.** Each row corresponds to a bacterial species or strain. Columns, from left to right, show species (strain) name, whether the null hypothesis of uniform distribution of the conditionally essential genes is rejected by Kuiper's test (1) or not (0), the P-value of the test, Kuiper's test statistics, and the critical value of this statistic above which the null hypothesis is rejected. In 52 among the 55 species the null hypothesis is rejected, i.e., conditionally essential genes are significantly clustered. We consider a metabolic gene conditionally essential, if its deletion abolishes viability on at least one carbon source. Note that clustering of conditionally essential genes may be of limited biological relevance, because different conditionally essential genes may not be essential on the same set of carbon sources.

| Species (strain) | (--) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| Methanosarcina barkeri str. Fusaro | 357 | 196 | 76 | 63 | 3.92E-02 | 1.5099 | 1 |
| Geobacter metallireducens GS-15 | 535 | 184 | 189 | 79 | 2.25E-01 | 1.2153 | 0 |
| *E. coli* BL21(DE3) [iB21 1397] | 932 | 241 | 106 | 58 | 5.31E-05 | 2.116 | 1 |
| *E. coli* BW2952 | 888 | 230 | 140 | 70 | 9.97E-05 | 1.9304 | 1 |
| *E. coli* CFT073 | 896 | 207 | 147 | 57 | 4.24E-03 | 1.6784 | 1 |
| *E. coli* O127:H6 | 898 | 226 | 105 | 55 | 9.29E-05 | 2.0813 | 1 |
| *E. coli* 042 | 903 | 252 | 94 | 65 | 5.54E-07 | 2.4778 | 1 |
| *E. coli* 55989 | 928 | 243 | 99 | 60 | 4.84E-06 | 2.3145 | 1 |
| *E. coli* ABU 83972 | 910 | 250 | 96 | 64 | 9.58E-07 | 2.4267 | 1 |
| *E. coli* B str. REL606 | 928 | 242 | 99 | 60 | 4.57E-06 | 2.3241 | 1 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 946 | 244 | 105 | 59 | 2.25E-05 | 2.1785 | 1 |
| *E. coli* BL21(DE3) [iECD1391] | 967 | 202 | 115 | 49 | 2.54E-04 | 2.0397 | 1 |
| *E. coli* DH1 [iEcDH1 1363] | 960 | 244 | 99 | 60 | 2.39E-06 | 2.3845 | 1 |
| *E. coli* O157:H7 | 872 | 231 | 98 | 61 | 4.51E-06 | 2.3497 | 1 |
| *E. coli* HS | 916 | 246 | 96 | 63 | 1.15E-06 | 2.4436 | 1 |
| *E. coli* NA114 | 897 | 246 | 99 | 59 | 3.30E-05 | 2.1731 | 1 |
| *E. coli* O103:H2 str. 12009 | 921 | 247 | 96 | 63 | 1.14E-06 | 2.447 | 1 |
| *E. coli* O111:H- str. 11128 | 919 | 250 | 100 | 59 | 3.40E-05 | 2.1688 | 1 |
| *E. coli* O26:H11 str. 11368 | 953 | 243 | 99 | 60 | 2.51E-06 | 2.3769 | 1 |
| *E. coli* IHE3034 | 914 | 230 | 99 | 61 | 1.34E-06 | 2.4486 | 1 |
| *E. coli* ATCC 8739 | 956 | 253 | 96 | 63 | 6.55E-07 | 2.4797 | 1 |
| *E. coli* 536 | 898 | 251 | 97 | 63 | 4.16E-06 | 2.3237 | 1 |
| *E. coli* O157:H7 str. Sakai | 913 | 229 | 100 | 59 | 5.44E-06 | 2.3523 | 1 |
| *E. coli* S88 | 920 | 225 | 100 | 60 | 1.20E-06 | 2.4533 | 1 |
| *E. coli* SE11 | 936 | 253 | 100 | 59 | 2.22E-05 | 2.1828 | 1 |
| *E. coli* SE15 | 894 | 223 | 134 | 76 | 8.92E-07 | 2.2737 | 1 |
| *E. coli* SMS-3-5 | 926 | 262 | 96 | 63 | 4.27E-06 | 2.3194 | 1 |
| *E. coli* O157:H7 str. TW14359 | 899 | 241 | 97 | 62 | 2.02E-06 | 2.3843 | 1 |
| *E. coli* W [iECW_1372] | 962 | 252 | 99 | 59 | 1.09E-05 | 2.2751 | 1 |
| *E. coli* KO11FL | 951 | 245 | 96 | 62 | 5.29E-07 | 2.5069 | 1 |
| *E. coli* ETEC H10407 | 917 | 257 | 99 | 60 | 2.61E-05 | 2.1625 | 1 |
| *E. coli* O55:H7 str. CB9615 | 869 | 255 | 99 | 60 | 7.39E-05 | 2.0654 | 1 |
| *E. coli* LF82 | 896 | 224 | 118 | 64 | 1.25E-05 | 2.1695 | 1 |
| *E. coli* O83:H1 str. NRG 857C | 918 | 233 | 103 | 57 | 2.70E-05 | 2.1803 | 1 |
| Shigella flexneri 2a str. 2457T | 784 | 237 | 104 | 63 | 1.12E-04 | 2.0039 | 1 |
| Shigella flexneri 5 str. 8401 | 763 | 257 | 97 | 67 | 6.48E-05 | 2.0507 | 1 |
| *E. coli* UM146 | 935 | 225 | 100 | 59 | 1.58E-06 | 2.4518 | 1 |
| *E. coli* UMNK88 | 919 | 275 | 96 | 63 | 1.47E-05 | 2.1931 | 1 |
| *E. coli* UTI89 | 915 | 235 | 98 | 62 | 9.38E-07 | 2.4633 | 1 |
| *E. coli* str. K-12 W3110 | 1141 | 58 | 130 | 29 | 2.48E-08 | 4.3885 | 1 |
| Klebsiella pneumoniae_MGH78578 | 844 | 296 | 60 | 29 | 1.72E-01 | 1.3782 | 0 |
| Bacillus subtilis str. 168 | 561 | 163 | 92 | 28 | 8.15E-01 | 1.0475 | 0 |
| Salmonella Typhimurium str. LT2 | 892 | 239 | 80 | 60 | 7.35E-08 | 2.7992 | 1 |

**Table S6: Horizontal gene transfer and strictly essential genes.** Each row corresponds to one of the 43 species strains for which information about horizontally transferred genes (HGT) is available in the HGTree database. Columns, from left to right, show the species (strain) name, the number of metabolic genes that are neither strictly essential nor HGT-acquired $(--)$, the number of metabolic genes that are not strictly essential, but HGT-acquired $(-+)$, the number of metabolic genes that are strictly essential, but are not HGT-acquired $(+-)$, and the number of metabolic genes that are both strictly essential and HGT-acquired $(++)$, the P value of a Fisher exact test on this data, the odds ratio (defined as the odds of being strictly essential among HGT-acquired genes divided by the odds of being strictly essential among non-HGT acquired essential genes), and whether the null hypothesis of a lack of association between a gene's strict essentiality and horizontal transfer is rejected (1) or not (0). In 40 of the 43 species (93.02%) the null hypothesis is rejected. Note that we consider a metabolic gene strictly essential, if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Species (strain) | (--) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| Methanosarcina barkeri str. Fusaro | 357 | 196 | 76 | 63 | 3.92E-02 | 1.5099 | 1 |
| Geobacter metallireducens GS-15 | 535 | 184 | 189 | 79 | 2.25E-01 | 1.2153 | 0 |
| *E. coli* BL21(DE3) [iB21 1397] | 808 | 203 | 230 | 96 | 5.68E-04 | 1.6613 | 1 |
| *E. coli* BW2952 | 764 | 197 | 264 | 103 | 4.15E-03 | 1.5131 | 1 |
| *E. coli* CFT073 | 896 | 207 | 147 | 57 | 4.24E-03 | 1.6784 | 1 |
| *E. coli* O127:H6 | 775 | 196 | 228 | 85 | 1.18E-02 | 1.4741 | 1 |
| *E. coli* 042 | 770 | 215 | 227 | 102 | 1.03E-03 | 1.6093 | 1 |
| *E. coli* 55989 | 799 | 205 | 228 | 98 | 4.55E-04 | 1.6753 | 1 |
| *E. coli* ABU 83972 | 779 | 221 | 227 | 93 | 1.27E-02 | 1.4441 | 1 |
| *E. coli* B str. REL606 | 799 | 203 | 228 | 99 | 2.55E-04 | 1.709 | 1 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 823 | 205 | 228 | 98 | 1.83E-04 | 1.7256 | 1 |
| *E. coli* BL21(DE3) [iECD1391] | 842 | 164 | 240 | 87 | 6.14E-05 | 1.8611 | 1 |
| *E. coli* DH1 [iEcDH1 1363] | 832 | 208 | 227 | 96 | 4.12E-04 | 1.6916 | 1 |
| *E. coli* O157:H7 | 712 | 193 | 258 | 99 | 1.76E-02 | 1.4156 | 1 |
| *E. coli* HS | 778 | 207 | 234 | 102 | 5.95E-04 | 1.6383 | 1 |
| *E. coli* NA114 | 768 | 213 | 228 | 92 | 1.21E-02 | 1.4549 | 1 |
| *E. coli* O103:H2 str. 12009 | 796 | 209 | 221 | 101 | 1.47E-04 | 1.7406 | 1 |
| *E. coli* O111:H- str. 11128 | 778 | 204 | 241 | 105 | 3.78E-04 | 1.6616 | 1 |
| *E. coli* O26:H11 str. 11368 | 813 | 205 | 239 | 98 | 8.81E-04 | 1.6262 | 1 |
| *E. coli* IHE3034 | 786 | 198 | 227 | 93 | 1.14E-03 | 1.6264 | 1 |
| *E. coli* ATCC 8739 | 830 | 215 | 222 | 101 | 1.11E-04 | 1.7563 | 1 |
| *E. coli* 536 | 767 | 222 | 228 | 92 | 2.39E-02 | 1.3941 | 1 |
| *E. coli* O157:H7 str. Sakai | 785 | 193 | 228 | 95 | 3.72E-04 | 1.6947 | 1 |
| *E. coli* S88 | 792 | 194 | 228 | 91 | 1.35E-03 | 1.6294 | 1 |
| *E. coli* SE11 | 810 | 215 | 226 | 97 | 1.11E-03 | 1.617 | 1 |
| *E. coli* SE15 | 763 | 189 | 265 | 110 | 2.59E-04 | 1.6758 | 1 |
| *E. coli* SMS-3-5 | 796 | 223 | 226 | 102 | 1.06E-03 | 1.611 | 1 |
| *E. coli* O157:H7 str. TW14359 | 771 | 206 | 225 | 97 | 1.08E-03 | 1.6135 | 1 |
| *E. coli* W [iECW_1372] | 832 | 214 | 229 | 97 | 6.36E-04 | 1.6468 | 1 |
| *E. coli* KO11FL | 825 | 208 | 222 | 99 | 9.54E-05 | 1.7688 | 1 |
| *E. coli* ETEC H10407 | 781 | 219 | 235 | 98 | 5.96E-03 | 1.4872 | 1 |
| *E. coli* O55:H7 str. CB9615 | 730 | 216 | 238 | 99 | 1.84E-02 | 1.4058 | 1 |
| *E. coli* LF82 | 712 | 187 | 302 | 101 | 9.66E-02 | 1.2734 | 0 |
| *E. coli* O83:H1 str. NRG 857C | 792 | 202 | 229 | 88 | 6.49E-03 | 1.5067 | 1 |
| Shigella flexneri 2a str. 2457T | 645 | 209 | 243 | 91 | 3.35E-01 | 1.1557 | 0 |
| Shigella flexneri 5 str. 8401 | 639 | 225 | 221 | 99 | 1.06E-01 | 1.2722 | 0 |
| *E. coli* UM146 | 806 | 194 | 229 | 90 | 1.30E-03 | 1.6328 | 1 |
| *E. coli* UMNK88 | 799 | 235 | 216 | 103 | 8.44E-04 | 1.6213 | 1 |
| *E. coli* UTI89 | 787 | 203 | 226 | 94 | 1.24E-03 | 1.6125 | 1 |
| *E. coli* str. K-12 W3110 | 981 | 48 | 290 | 39 | 1.36E-05 | 2.7485 | 1 |
| Klebsiella pneumoniae_MGH78578 | 608 | 219 | 296 | 106 | 1.00E+00 | 0.9942 | 0 |
| Bacillus subtilis str. 168 | 425 | 131 | 228 | 60 | 3.87E-01 | 0.8538 | 0 |
| Salmonella Typhimurium str. LT2 | 773 | 205 | 199 | 94 | 1.17E-04 | 1.7811 | 1 |

**Table S7: Horizontal gene transfer and conditionally essential genes.** Each row corresponds to one of the 43 species (strain) for which information about horizontally transferred genes (HGT) is available in the HGTree database. Columns, from left to right, show the species (strain) name, the number of metabolic genes that are neither conditionally essential nor HGT-acquired ($--$), the number of metabolic genes that are not conditionally essential, but HGT-acquired ($-+$), the number of metabolic genes that are conditionally essential, but are not HGT-acquired ($+-$), and the number of metabolic genes that are both conditionally essential and HGT-acquired ($++$), the P value of a Fisher exact test on this data, the odds ratio (defined as the odds of being conditionally essential among HGT acquired metabolic genes divided by the odds of being conditionally essential among non-HGT acquired metabolic genes), and whether the null hypothesis of a lack of association between a gene's conditional essentiality and horizontal transfer is rejected (1) or not (0). In 37 of the 43 species (86.05%) the null hypothesis is rejected. Note that we consider a metabolic gene conditionally essential, if its deletion abolishes viability on at least one carbon source

| | HGT-acquired essential genes | | | | Non-HGT acquired essential genes | | | | All essential genes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Species(strain) | Hypothesis rejected | P-value | Kuiper test statistic | Critical value | Hypothesis rejected | P-value | Kuiper test statistic | Critical value | Hypothesis rejected | P-value | Kuiper test statistic | Critical value |
| Methanosarcina barkeri str. Fusaro | 0 | 7.87E-01 | 0.1285 | 0.2182 | 0 | 3.30E-01 | 0.1556 | 0.2011 | 0 | 2.53E-01 | 0.1211 | 0.1492 |
| Geobacter metallireducens GS-15 | 0 | 2.99E-01 | 0.1517 | 0.1922 | 1 | 1.72E-05 | 0.2029 | 0.1293 | 1 | 1.84E-05 | 0.1688 | 0.1078 |
| *E. coli* BL21(DE3) [iB21 1397] | 1 | 9.14E-03 | 0.2581 | 0.2236 | 0 | 2.51E-01 | 0.1368 | 0.1684 | 1 | 3.76E-03 | 0.1656 | 0.1354 |
| *E. coli* BW2952 | 1 | 1.51E-03 | 0.2635 | 0.204 | 1 | 3.50E-02 | 0.1516 | 0.1465 | 1 | 2.31E-04 | 0.1694 | 0.1197 |
| *E. coli* CFT073 | 1 | 6.02E-04 | 0.3108 | 0.2295 | 1 | 1.59E-02 | 0.1621 | 0.1465 | 1 | 5.52E-04 | 0.1687 | 0.1242 |
| *E. coli* O127:H6 | 1 | 4.04E-03 | 0.2796 | 0.2295 | 0 | 1.26E-01 | 0.1524 | 0.1692 | 1 | 2.15E-03 | 0.1733 | 0.1371 |
| *E. coli* 042 | 1 | 7.24E-03 | 0.2482 | 0.2114 | 1 | 2.95E-02 | 0.1879 | 0.1786 | 1 | 7.55E-04 | 0.1839 | 0.1375 |
| *E. coli* 55989 | 1 | 2.71E-03 | 0.2747 | 0.22 | 1 | 2.98E-02 | 0.1829 | 0.1744 | 1 | 6.16E-04 | 0.1858 | 0.1375 |
| *E. coli* ABU 83972 | 1 | 5.13E-03 | 0.2559 | 0.213 | 0 | 7.41E-02 | 0.1698 | 0.1769 | 1 | 1.17E-03 | 0.1792 | 0.1371 |
| *E. coli* B str. REL606 | 1 | 6.45E-03 | 0.2601 | 0.22 | 0 | 8.18E-02 | 0.1654 | 0.1744 | 1 | 3.72E-03 | 0.1682 | 0.1375 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 1 | 2.32E-03 | 0.2794 | 0.2218 | 0 | 3.43E-01 | 0.1296 | 0.1692 | 1 | 6.43E-03 | 0.1599 | 0.1354 |
| *E. coli* BL21(DE3) [iECD1391] | 1 | 7.69E-03 | 0.2835 | 0.2432 | 0 | 1.41E-01 | 0.1434 | 0.1617 | 1 | 2.63E-03 | 0.1692 | 0.1354 |
| *E. coli* DH1 [iEcDH1 1363] | 1 | 2.31E-03 | 0.2773 | 0.22 | 1 | 2.55E-02 | 0.1854 | 0.1744 | 1 | 6.30E-04 | 0.1856 | 0.1375 |
| *E. coli* O157:H7 | 1 | 1.75E-02 | 0.2399 | 0.2182 | 1 | 1.50E-02 | 0.1946 | 0.1752 | 1 | 9.67E-04 | 0.1816 | 0.1375 |
| *E. coli* HS | 1 | 1.31E-02 | 0.2415 | 0.2147 | 1 | 2.32E-02 | 0.1898 | 0.1769 | 1 | 1.74E-03 | 0.1759 | 0.1375 |
| *E. coli* NA114 | 1 | 1.17E-02 | 0.2514 | 0.2218 | 1 | 1.11E-02 | 0.1981 | 0.1744 | 1 | 2.84E-03 | 0.1716 | 0.138 |
| *E. coli* O103:H2 str. 12009 | 1 | 1.62E-02 | 0.2376 | 0.2147 | 1 | 2.00E-02 | 0.1921 | 0.1769 | 1 | 1.37E-03 | 0.1782 | 0.1375 |
| *E. coli* O111:H- str. 11128 | 1 | 2.82E-03 | 0.2763 | 0.2218 | 1 | 1.98E-02 | 0.1884 | 0.1736 | 1 | 8.53E-04 | 0.1828 | 0.1375 |
| *E. coli* O26:H11 str. 11368 | 1 | 1.16E-02 | 0.2496 | 0.22 | 1 | 3.48E-02 | 0.1804 | 0.1744 | 1 | 8.09E-04 | 0.1833 | 0.1375 |
| *E. coli* IHE3034 | 1 | 3.41E-03 | 0.2687 | 0.2182 | 0 | 1.65E-01 | 0.1512 | 0.1744 | 1 | 2.31E-03 | 0.1726 | 0.1371 |
| *E. coli* ATCC 8739 | 1 | 2.40E-03 | 0.2702 | 0.2147 | 0 | 6.37E-02 | 0.1726 | 0.1769 | 1 | 1.15E-03 | 0.18 | 0.1375 |
| *E. coli* 536 | 1 | 1.73E-02 | 0.2363 | 0.2147 | 1 | 4.72E-02 | 0.177 | 0.176 | 1 | 1.31E-03 | 0.1782 | 0.1371 |
| *E. coli* O157:H7 str. Sakai | 1 | 1.00E-03 | 0.2927 | 0.2218 | 0 | 8.97E-02 | 0.1628 | 0.1736 | 1 | 1.00E-03 | 0.1812 | 0.1375 |
| *E. coli* S88 | 1 | 6.01E-03 | 0.2613 | 0.22 | 0 | 6.71E-02 | 0.1682 | 0.1736 | 1 | 2.07E-03 | 0.1737 | 0.1371 |
| *E. coli* SE11 | 1 | 5.72E-03 | 0.2643 | 0.2218 | 1 | 2.15E-02 | 0.1871 | 0.1736 | 1 | 6.58E-04 | 0.1852 | 0.1375 |
| *E. coli* SE15 | 1 | 3.79E-04 | 0.2715 | 0.1958 | 0 | 1.46E-01 | 0.1322 | 0.1498 | 1 | 1.51E-03 | 0.1543 | 0.1197 |
| *E. coli* SMS-3-5 | 1 | 4.64E-03 | 0.2595 | 0.2147 | 0 | 9.30E-02 | 0.1654 | 0.1769 | 1 | 7.94E-04 | 0.1835 | 0.1375 |
| *E. coli* O157:H7 str. TW14359 | 1 | 9.58E-03 | 0.249 | 0.2164 | 1 | 4.57E-02 | 0.1776 | 0.176 | 1 | 1.15E-03 | 0.1799 | 0.1375 |
| *E. coli* W [iECW_1372] | 1 | 1.39E-03 | 0.2876 | 0.2218 | 1 | 1.93E-02 | 0.1898 | 0.1744 | 1 | 2.09E-04 | 0.196 | 0.138 |
| *E. coli* KO11FL | 1 | 6.65E-03 | 0.2554 | 0.2164 | 1 | 1.02E-02 | 0.2024 | 0.1769 | 1 | 5.37E-04 | 0.1876 | 0.138 |
| *E. coli* ETEC H10407 | 1 | 3.00E-03 | 0.273 | 0.22 | 1 | 3.26E-02 | 0.1814 | 0.1744 | 1 | 7.23E-04 | 0.1843 | 0.1375 |
| *E. coli* O55:H7 str. CB9615 | 1 | 7.71E-03 | 0.2569 | 0.22 | 1 | 2.02E-02 | 0.1891 | 0.1744 | 1 | 3.12E-04 | 0.192 | 0.1375 |
| *E. coli* LF82 | 1 | 1.40E-05 | 0.3511 | 0.2218 | 1 | 6.74E-04 | 0.232 | 0.1728 | 1 | 2.06E-09 | 0.275 | 0.1371 |
| *E. coli* O83:H1 str. NRG 857C | 1 | 4.10E-03 | 0.2746 | 0.2255 | 0 | 7.09E-02 | 0.1647 | 0.1708 | 1 | 1.59E-03 | 0.1763 | 0.1371 |
| Shigella flexneri 2a str. 2457T | 1 | 7.07E-03 | 0.2524 | 0.2147 | 0 | 1.54E-01 | 0.1505 | 0.172 | 1 | 1.13E-02 | 0.1533 | 0.135 |
| Shigella flexneri 5 str. 8401 | 1 | 1.71E-02 | 0.2295 | 0.2083 | 0 | 1.85E-01 | 0.1502 | 0.176 | 1 | 9.77E-03 | 0.1554 | 0.1354 |
| *E. coli* UM146 | 1 | 6.30E-03 | 0.2626 | 0.2218 | 0 | 6.40E-02 | 0.169 | 0.1736 | 1 | 7.51E-04 | 0.184 | 0.1375 |
| *E. coli* UMNK88 | 0 | 4.37E-01 | 0.156 | 0.2147 | 0 | 2.20E-01 | 0.1469 | 0.1769 | 0 | 1.73E-01 | 0.1185 | 0.1375 |
| *E. coli* UTI89 | 1 | 2.38E-03 | 0.2724 | 0.2164 | 0 | 8.33E-02 | 0.1659 | 0.1752 | 1 | 1.30E-03 | 0.1782 | 0.1371 |
| *E. coli* str. K-12 W3110 | 1 | 2.26E-03 | 0.3942 | 0.3115 | 1 | 2.87E-02 | 0.1602 | 0.1521 | 1 | 9.73E-04 | 0.1815 | 0.1375 |
| Klebsiella pneumoniae_MGH78578 | 1 | 1.15E-02 | 0.3546 | 0.3115 | 0 | 2.98E-01 | 0.1842 | 0.2339 | 1 | 4.24E-02 | 0.1922 | 0.1887 |
| Bacillus subtilis str. 168 | 0 | 2.27E-01 | 0.2629 | 0.3163 | 1 | 1.99E-04 | 0.2574 | 0.1804 | 1 | 4.12E-04 | 0.218 | 0.1583 |
| Salmonella Typhimurium str. LT2 | 1 | 6.13E-04 | 0.2977 | 0.22 | 1 | 4.30E-02 | 0.1967 | 0.1933 | 1 | 3.02E-04 | 0.2048 | 0.1465 |

**Table S8: Horizontal gene transfer plays an active role in the clustering of the strictly essential genes.** In this analysis we divided the set of strictly non-essential genes into two groups: *i)* those acquired via horizontal gene transfer, and *ii)* those not acquired through horizontal gene transfer. Using Kuiper's test, we examined the clustering of *i)* the first group of strictly essential genes alone (columns 2-5; labeled red in the first row), ii) the second group of strictly essential genes alone (columns 6-9; labeled blue in the first row), and *iii)* all strictly essential genes together (columns 10-13; labeled black in the first row). Each row corresponds to a bacterial species (strain). The first column is the species or strain name, and in each of the three sets of four columns, from left to right, columns show whether the null hypothesis of uniform distribution of strictly essential genes is rejected by Kuiper's test (1) or not (0), the P-value of the test, Kuiper's test statistics, and the critical value of this statistic above which the null hypothesis is rejected. In 18 out of the 43 genomes (41.86%, colored blue) used in this analysis, the strictly essential genes acquired by horizontal gene transfer are significantly clustered (first group), but the strictly essential genes not acquired by horizontal gene transfer (second group) are not significantly clustered. Only in two species (4.65%, colored red) are the strictly essential genes of the second group significantly clustered but genes in the first group are not. In two other species neither of the groups of strictly essential genes are significantly clustered (4.65%, colored green). Finally, in 21 out of the 43 genomes (48.83%, colored black) the strictly essential genes of both groups are significantly clustered. Note that we consider a metabolic gene strictly essential, if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Species (strains) | Number of operons | Average number of genes per operon | Number of operons with at least one conditionally essential gene | Fraction of operons with at least one conditionally essential gene | Number of operons with more than one conditionally essential gene | Fraction of operons with more than one conditionally essential gene | Number of operons with at least one strictly essential gene | Fraction of operons with at least one strictly essential gene | Number of operons with more than one strictly essential gene | Fraction of operons with more than one strictly essential gene |
|---|---|---|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 241 | 3.07 | 95 | 0.39 | 62 | 0.26 | 48 | 0.2 | 25 | 0.1 |
| Methanosarcina barkeri str. Fusaro | 91 | 2.75 | 20 | 0.22 | 15 | 0.16 | 20 | 0.22 | 15 | 0.16 |
| Geobacter metallireducens GS-15 | 203 | 3.12 | 67 | 0.33 | 47 | 0.23 | 67 | 0.33 | 47 | 0.23 |
| *E. coli* APEC O1 | 247 | 2.94 | 101 | 0.41 | 59 | 0.24 | 52 | 0.21 | 29 | 0.12 |
| *E. coli* BL21(DE3) [iB21 1397] | 277 | 3.19 | 108 | 0.39 | 63 | 0.23 | 59 | 0.21 | 31 | 0.11 |
| *E. coli* BW2952 | 247 | 3.14 | 113 | 0.46 | 71 | 0.29 | 69 | 0.28 | 41 | 0.17 |
| *E. coli* CFT073 | 260 | 2.8 | 61 | 0.23 | 31 | 0.12 | 61 | 0.23 | 31 | 0.12 |
| *E. coli* O127:H6 | 244 | 3.01 | 97 | 0.4 | 59 | 0.24 | 53 | 0.22 | 29 | 0.12 |
| *E. coli* 042 | 252 | 3.07 | 100 | 0.4 | 64 | 0.25 | 54 | 0.21 | 30 | 0.12 |
| *E. coli* 55989 | 257 | 3.05 | 101 | 0.39 | 61 | 0.24 | 55 | 0.21 | 29 | 0.11 |
| *E. coli* ABU 83972 | 247 | 3.07 | 99 | 0.4 | 61 | 0.25 | 51 | 0.21 | 29 | 0.12 |
| *E. coli* B str. REL606 | 256 | 3 | 99 | 0.39 | 62 | 0.24 | 55 | 0.21 | 29 | 0.11 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 256 | 3.05 | 98 | 0.38 | 62 | 0.24 | 56 | 0.22 | 31 | 0.12 |
| *E. coli* BL21(DE3) [iECD1391] | 253 | 3.02 | 99 | 0.39 | 63 | 0.25 | 56 | 0.22 | 31 | 0.12 |
| *E. coli* DH1 [iEcDH1 1363] | 260 | 3.09 | 101 | 0.39 | 60 | 0.23 | 56 | 0.22 | 29 | 0.11 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 255 | 3.1 | 101 | 0.4 | 60 | 0.24 | 56 | 0.22 | 29 | 0.11 |
| *E. coli* E24377A | 260 | 2.98 | 99 | 0.38 | 61 | 0.23 | 54 | 0.21 | 29 | 0.11 |
| *E. coli* ED1a | 242 | 3 | 103 | 0.43 | 57 | 0.24 | 53 | 0.22 | 28 | 0.12 |
| *E. coli* O157:H7 | 240 | 2.95 | 98 | 0.41 | 58 | 0.24 | 54 | 0.23 | 28 | 0.12 |
| *E. coli* HS | 253 | 3.06 | 100 | 0.4 | 62 | 0.25 | 55 | 0.22 | 29 | 0.11 |
| *E. coli* NA114 | 245 | 3.03 | 98 | 0.4 | 60 | 0.24 | 52 | 0.21 | 29 | 0.12 |
| *E. coli* O103:H2 str. 12009 | 250 | 3.06 | 99 | 0.4 | 59 | 0.24 | 55 | 0.22 | 29 | 0.12 |
| *E. coli* O111:H- str. 11128 | 251 | 3.05 | 97 | 0.39 | 57 | 0.23 | 55 | 0.22 | 29 | 0.12 |
| *E. coli* O26:H11 str. 11368 | 261 | 3.04 | 100 | 0.38 | 58 | 0.22 | 55 | 0.21 | 29 | 0.11 |
| *E. coli* IHE3034 | 242 | 3.01 | 99 | 0.41 | 59 | 0.24 | 51 | 0.21 | 28 | 0.12 |
| *E. coli* ATCC 8739 | 258 | 3.1 | 98 | 0.38 | 61 | 0.24 | 54 | 0.21 | 29 | 0.11 |
| *E. coli* 536 | 248 | 2.97 | 100 | 0.4 | 61 | 0.25 | 52 | 0.21 | 29 | 0.12 |
| *E. coli* O157:H7 str. Sakai | 248 | 2.99 | 98 | 0.4 | 58 | 0.23 | 54 | 0.22 | 28 | 0.11 |
| *E. coli* S88 | 248 | 3.01 | 99 | 0.4 | 60 | 0.24 | 51 | 0.21 | 29 | 0.12 |
| *E. coli* SE11 | 259 | 3.08 | 100 | 0.39 | 60 | 0.23 | 55 | 0.21 | 29 | 0.11 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *E. coli* SE15 | 251 | 3.07 | 117 | 0.47 | 73 | 0.29 | 67 | 0.27 | 41 | 0.16 |
| *E. coli* SMS-3-5 | 260 | 3.08 | 103 | 0.4 | 66 | 0.25 | 55 | 0.21 | 30 | 0.12 |
| *E. coli* O157:H7 str. TW14359 | 245 | 3 | 99 | 0.4 | 59 | 0.24 | 55 | 0.22 | 29 | 0.12 |
| *E. coli* UMN026 | 259 | 3.05 | 100 | 0.39 | 64 | 0.25 | 54 | 0.21 | 30 | 0.12 |
| *E. coli* W [iECW_1372] | 259 | 3.03 | 100 | 0.39 | 61 | 0.24 | 56 | 0.22 | 29 | 0.11 |
| *E. coli* KO11FL | 258 | 3.09 | 98 | 0.38 | 61 | 0.24 | 54 | 0.21 | 29 | 0.11 |
| *E. coli* ETEC H10407 | 257 | 3.01 | 100 | 0.39 | 60 | 0.23 | 55 | 0.21 | 29 | 0.11 |
| *E. coli* O55:H7 str. CB9615 | 247 | 3.02 | 100 | 0.4 | 61 | 0.25 | 54 | 0.22 | 28 | 0.11 |
| *E. coli* K-12 MG1655 [iJO1366] | 262 | 3.12 | 115 | 0.44 | 75 | 0.29 | 68 | 0.26 | 41 | 0.16 |
| *E. coli* K-12 MG1655 [iJR904] | 181 | 2.87 | 81 | 0.45 | 58 | 0.32 | 35 | 0.19 | 21 | 0.12 |
| *E. coli* LF82 | 246 | 3.02 | 98 | 0.4 | 60 | 0.24 | 53 | 0.22 | 28 | 0.11 |
| Mycobacterium tuberculosis H37Rv | 122 | 2.82 | 55 | 0.45 | 30 | 0.25 | 55 | 0.45 | 30 | 0.25 |
| *E. coli* O83:H1 str. NRG 857C | 248 | 3.03 | 99 | 0.4 | 61 | 0.25 | 53 | 0.21 | 29 | 0.12 |
| Shigella flexneri 2a str. 2457T | 218 | 2.93 | 90 | 0.41 | 55 | 0.25 | 57 | 0.26 | 31 | 0.14 |
| Shigella flexneri 5 str. 8401 | 218 | 2.94 | 86 | 0.39 | 52 | 0.24 | 54 | 0.25 | 30 | 0.14 |
| *E. coli* UM146 | 249 | 3 | 99 | 0.4 | 61 | 0.24 | 51 | 0.2 | 29 | 0.12 |
| *E. coli* UMNK88 | 259 | 3.03 | 98 | 0.38 | 60 | 0.23 | 55 | 0.21 | 29 | 0.11 |
| *E. coli* UTI89 | 259 | 2.95 | 101 | 0.39 | 60 | 0.23 | 53 | 0.2 | 29 | 0.11 |
| Klebsiella pneumoniae MGH78578 | 227 | 3 | 70 | 0.31 | 44 | 0.19 | 24 | 0.11 | 10 | 0.04 |
| Bacillus subtilis str. 168 | 161 | 3.24 | 76 | 0.47 | 52 | 0.32 | 33 | 0.2 | 25 | 0.16 |
| *E. coli* O157:H7 str. EDL933 | 253 | 3.02 | 98 | 0.39 | 59 | 0.23 | 54 | 0.21 | 28 | 0.11 |
| Salmonella Typhimurium str. LT2 | 247 | 2.95 | 82 | 0.33 | 48 | 0.19 | 43 | 0.17 | 23 | 0.09 |

**Table S9: Metabolic genes in Operons.** Each row corresponds to a given species (strain). Columns, from left to right, show species or strain names, the total number of operons in the genome, the average number of genes per operon, the number of operons with at least one conditionally essential gene, the fraction of operons with at least one conditionally essential gene, the number of operons with more than one conditionally essential metabolic gene, the fraction of operons with more than one conditionally essential metabolic gene, the number of operons with at least one strictly essential gene, the fraction of operons with at least one strictly essential gene, the number of operons with more than one strictly essential metabolic gene, and the fraction of operons with more than one strictly essential metabolic gene.

| Species (strain) | (--) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 470 | 647 | 51 | 93 | 1.50E-01 | 1.3247 | 0 |
| Methanosarcina barkeri str. Fusaro | 348 | 211 | 94 | 39 | 7.15E-02 | 0.6843 | 0 |
| Geobacter metallireducens GS-15 | 271 | 459 | 83 | 174 | 1.74E-01 | 1.2377 | 0 |
| *E. coli* APEC O1 | 528 | 627 | 60 | 98 | 7.33E-02 | 1.3754 | 0 |
| *E. coli* BL21(DE3) [iB21 1397] | 415 | 760 | 39 | 123 | 4.54E-03 | 1.7222 | 1 |
| *E. coli* BW2952 | 490 | 630 | 63 | 145 | 3.05E-04 | 1.7901 | 1 |
| *E. coli* CFT073 | 498 | 616 | 80 | 113 | 4.33E-01 | 1.1419 | 0 |
| *E. coli* O127:H6 | 497 | 629 | 52 | 106 | 7.74E-03 | 1.6107 | 1 |
| *E. coli* 042 | 488 | 669 | 52 | 105 | 3.09E-02 | 1.4729 | 1 |
| *E. coli* 55989 | 497 | 676 | 49 | 108 | 7.43E-03 | 1.6205 | 1 |
| *E. coli* ABU 83972 | 508 | 654 | 53 | 105 | 1.63E-02 | 1.5389 | 1 |
| *E. coli* B str. REL606 | 511 | 661 | 49 | 108 | 3.35E-03 | 1.7039 | 1 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 525 | 667 | 49 | 113 | 9.11E-04 | 1.8152 | 1 |
| *E. coli* BL21(DE3) [iECD1391] | 518 | 653 | 50 | 112 | 1.25E-03 | 1.7769 | 1 |
| *E. coli* DH1 [iEcDH1 1363] | 511 | 695 | 49 | 108 | 7.50E-03 | 1.6206 | 1 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 602 | 681 | 46 | 110 | 3.81E-05 | 2.1139 | 1 |
| *E. coli* E24377A | 516 | 668 | 50 | 107 | 5.80E-03 | 1.6531 | 1 |
| *E. coli* ED1a | 499 | 621 | 55 | 104 | 2.07E-02 | 1.5194 | 1 |
| *E. coli* O157:H7 | 502 | 603 | 51 | 106 | 2.54E-03 | 1.7303 | 1 |
| *E. coli* HS | 498 | 666 | 49 | 108 | 5.72E-03 | 1.6481 | 1 |
| *E. coli* NA114 | 504 | 641 | 54 | 102 | 3.09E-02 | 1.4852 | 1 |
| *E. coli* O103:H2 str. 12009 | 512 | 658 | 50 | 107 | 4.51E-03 | 1.6652 | 1 |
| *E. coli* O111:H- str. 11128 | 513 | 658 | 49 | 108 | 2.59E-03 | 1.7184 | 1 |
| *E. coli* O26:H11 str. 11368 | 512 | 686 | 49 | 108 | 5.80E-03 | 1.645 | 1 |
| *E. coli* IHE3034 | 521 | 625 | 54 | 104 | 7.97E-03 | 1.6055 | 1 |
| *E. coli* ATCC 8739 | 517 | 694 | 50 | 107 | 9.75E-03 | 1.5942 | 1 |
| *E. coli* 536 | 518 | 633 | 54 | 104 | 1.03E-02 | 1.576 | 1 |
| *E. coli* O157:H7 str. Sakai | 509 | 635 | 51 | 106 | 4.51E-03 | 1.666 | 1 |
| *E. coli* S88 | 506 | 641 | 53 | 105 | 1.28E-02 | 1.5639 | 1 |
| *E. coli* SE11 | 502 | 689 | 49 | 108 | 9.46E-03 | 1.6059 | 1 |
| *E. coli* SE15 | 492 | 627 | 64 | 144 | 4.14E-04 | 1.7656 | 1 |
| *E. coli* SMS-3-5 | 497 | 693 | 48 | 109 | 7.29E-03 | 1.6286 | 1 |
| *E. coli* O157:H7 str. TW14359 | 513 | 629 | 52 | 105 | 5.89E-03 | 1.6468 | 1 |
| *E. coli* UMN026 | 491 | 683 | 50 | 108 | 1.56E-02 | 1.5528 | 1 |
| *E. coli* W [iECW_1372] | 539 | 677 | 47 | 109 | 7.64E-04 | 1.8464 | 1 |
| *E. coli* KO11FL | 509 | 689 | 49 | 107 | 9.32E-03 | 1.6132 | 1 |
| *E. coli* ETEC H10407 | 511 | 665 | 49 | 108 | 3.41E-03 | 1.6937 | 1 |
| *E. coli* O55:H7 str. CB9615 | 485 | 641 | 51 | 106 | 1.22E-02 | 1.5726 | 1 |
| *E. coli* K-12 MG1655 [iJO1366] | 487 | 673 | 62 | 145 | 1.19E-03 | 1.6924 | 1 |
| *E. coli* K-12 MG1655 [iJR904] | 329 | 443 | 55 | 77 | 8.50E-01 | 1.0397 | 0 |
| *E. coli* LF82 | 507 | 637 | 53 | 105 | 1.02E-02 | 1.5768 | 1 |
| Mycobacterium tuberculosis H37Rv | 233 | 231 | 84 | 113 | 8.86E-02 | 1.3569 | 0 |
| *E. coli* O83:H1 str. NRG 857C | 507 | 646 | 52 | 106 | 9.91E-03 | 1.5998 | 1 |
| Shigella flexneri 2a str. 2457T | 495 | 530 | 54 | 109 | 3.60E-04 | 1.8852 | 1 |
| Shigella flexneri 5 str. 8401 | 490 | 532 | 54 | 108 | 4.97E-04 | 1.8421 | 1 |
| *E. coli* UM146 | 519 | 643 | 52 | 105 | 6.05E-03 | 1.6298 | 1 |
| *E. coli* UMNK88 | 518 | 678 | 49 | 108 | 4.41E-03 | 1.6839 | 1 |
| *E. coli* UTI89 | 494 | 658 | 53 | 105 | 2.56E-02 | 1.4874 | 1 |
| Klebsiella pneumoniae_MGH78578 | 514 | 633 | 35 | 47 | 7.32E-01 | 1.0904 | 0 |
| Bacillus subtilis str. 168 | 285 | 441 | 37 | 81 | 1.03E-01 | 1.4148 | 0 |
| *E. coli* O157:H7 str. EDL933 | 494 | 657 | 51 | 106 | 1.55E-02 | 1.5628 | 1 |
| Salmonella Typhimurium str. LT2 | 483 | 650 | 59 | 79 | 1.00E+00 | 0.995 | 0 |

**Table S10: Operons and the strictly essential genes.** Each row corresponds to a given species or strain. Columns, from left to right, show species (strain) names, the number of metabolic genes that are neither strictly essential nor belong to an operon $(--)$, the number of metabolic genes that are not strictly essential but do belong to an operon $(-+)$, the number of metabolic genes that are strictly essential, but do not belong to an operon $(+-)$, the number of metabolic genes that are both strictly essential and belong to an operon $(++)$, the P value of a Fisher's exact test on this data, the odds ratio (defined as the odds of being strictly essential for operonic metabolic genes divided by the odds of being strictly essential for non-operonic metabolic genes), and whether the null hypothesis of a lack of association between a gene's strict essentiality and being part of an operon is rejected (1) or not (0). In 42 of the 52 species (80.76%) the null hypothesis is rejected. Note that we consider a metabolic gene strictly essential, if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Strain (species) | (--) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 411 | 533 | 110 | 207 | 5.67E-03 | 1.4511 | 1 |
| Methanosarcina barkeri str. Fusaro | 348 | 211 | 94 | 39 | 7.15E-02 | 0.6843 | 0 |
| Geobacter metallireducens GS-15 | 271 | 459 | 83 | 174 | 1.74E-01 | 1.2377 | 0 |
| *E. coli* APEC O1 | 474 | 530 | 114 | 195 | 1.66E-03 | 1.5298 | 1 |
| *E. coli* BL21(DE3) [iB21 1397] | 377 | 644 | 77 | 239 | 3.23E-05 | 1.817 | 1 |
| *E. coli* BW2952 | 437 | 536 | 116 | 239 | 5.68E-05 | 1.6798 | 1 |
| *E. coli* CFT073 | 498 | 616 | 80 | 113 | 4.33E-01 | 1.1419 | 0 |
| *E. coli* O127:H6 | 447 | 535 | 102 | 200 | 3.22E-04 | 1.6383 | 1 |
| *E. coli* 042 | 439 | 558 | 101 | 216 | 1.39E-04 | 1.6825 | 1 |
| *E. coli* 55989 | 442 | 571 | 104 | 213 | 6.59E-04 | 1.5854 | 1 |
| *E. coli* ABU 83972 | 459 | 552 | 102 | 207 | 1.32E-04 | 1.6875 | 1 |
| *E. coli* B str. REL606 | 456 | 557 | 104 | 212 | 1.50E-04 | 1.6688 | 1 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 471 | 567 | 103 | 213 | 5.43E-05 | 1.7178 | 1 |
| *E. coli* BL21(DE3) [iECD1391] | 465 | 551 | 103 | 214 | 3.01E-05 | 1.7534 | 1 |
| *E. coli* DH1 [iEcDH1 1363] | 456 | 594 | 104 | 209 | 1.33E-03 | 1.5427 | 1 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 548 | 580 | 100 | 211 | 2.34E-07 | 1.9936 | 1 |
| *E. coli* E24377A | 464 | 566 | 102 | 209 | 1.40E-04 | 1.6798 | 1 |
| *E. coli* ED1a | 453 | 516 | 101 | 209 | 1.31E-05 | 1.8167 | 1 |
| *E. coli* O157:H7 | 449 | 500 | 104 | 209 | 1.38E-05 | 1.8046 | 1 |
| *E. coli* HS | 445 | 559 | 102 | 215 | 1.44E-04 | 1.678 | 1 |
| *E. coli* NA114 | 451 | 540 | 107 | 203 | 6.28E-04 | 1.5845 | 1 |
| *E. coli* O103:H2 str. 12009 | 461 | 559 | 101 | 206 | 1.30E-04 | 1.682 | 1 |
| *E. coli* O111:H- str. 11128 | 461 | 560 | 101 | 206 | 1.31E-04 | 1.679 | 1 |
| *E. coli* O26:H11 str. 11368 | 456 | 586 | 105 | 208 | 1.33E-03 | 1.5415 | 1 |
| *E. coli* IHE3034 | 468 | 527 | 107 | 202 | 1.38E-04 | 1.6765 | 1 |
| *E. coli* ATCC 8739 | 463 | 592 | 104 | 209 | 8.47E-04 | 1.5717 | 1 |
| *E. coli* 536 | 469 | 531 | 103 | 206 | 2.58E-05 | 1.7665 | 1 |
| *E. coli* O157:H7 str. Sakai | 456 | 538 | 104 | 203 | 2.17E-04 | 1.6544 | 1 |
| *E. coli* S88 | 456 | 541 | 103 | 205 | 1.32E-04 | 1.6776 | 1 |
| *E. coli* SE11 | 450 | 586 | 101 | 211 | 4.92E-04 | 1.6043 | 1 |
| *E. coli* SE15 | 442 | 520 | 114 | 251 | 1.11E-06 | 1.8715 | 1 |
| *E. coli* SMS-3-5 | 449 | 580 | 96 | 222 | 2.04E-05 | 1.7902 | 1 |
| *E. coli* O157:H7 str. TW14359 | 459 | 531 | 106 | 203 | 1.79E-04 | 1.6554 | 1 |
| *E. coli* UMN026 | 438 | 573 | 103 | 218 | 3.33E-04 | 1.6179 | 1 |
| *E. coli* W [iECW_1372] | 487 | 575 | 99 | 211 | 1.18E-05 | 1.8051 | 1 |
| *E. coli* KO11FL | 456 | 586 | 102 | 210 | 5.03E-04 | 1.6021 | 1 |
| *E. coli* ETEC H10407 | 462 | 562 | 98 | 211 | 2.53E-05 | 1.77 | 1 |
| *E. coli* O55:H7 str. CB9615 | 435 | 527 | 101 | 220 | 1.53E-05 | 1.798 | 1 |
| *E. coli* K-12 MG1655 [iJO1366] | 434 | 562 | 115 | 256 | 2.38E-05 | 1.7191 | 1 |
| *E. coli* K-12 MG1655 [iJR904] | 262 | 331 | 122 | 189 | 1.57E-01 | 1.2262 | 0 |
| *E. coli* LF82 | 459 | 537 | 101 | 205 | 5.45E-05 | 1.7349 | 1 |
| Mycobacterium tuberculosis H37Rv | 233 | 231 | 84 | 113 | 8.86E-02 | 1.3569 | 0 |
| *E. coli* O83:H1 str. NRG 857C | 458 | 547 | 101 | 205 | 9.63E-05 | 1.6995 | 1 |
| Shigella flexneri 2a str. 2457T | 441 | 454 | 108 | 185 | 2.59E-04 | 1.6639 | 1 |
| Shigella flexneri 5 str. 8401 | 437 | 462 | 107 | 178 | 1.07E-03 | 1.5735 | 1 |
| *E. coli* UM146 | 469 | 542 | 102 | 206 | 3.47E-05 | 1.7476 | 1 |
| *E. coli* UMNK88 | 466 | 579 | 101 | 207 | 2.29E-04 | 1.6495 | 1 |
| *E. coli* UTI89 | 442 | 559 | 105 | 204 | 1.53E-03 | 1.5362 | 1 |
| Klebsiella pneumoniae_MGH78578 | 452 | 528 | 97 | 152 | 4.57E-02 | 1.3415 | 1 |
| Bacillus subtilis str. 168 | 246 | 336 | 76 | 186 | 2.35E-04 | 1.7918 | 1 |
| *E. coli* O157:H7 str. EDL933 | 443 | 558 | 102 | 205 | 5.80E-04 | 1.5956 | 1 |
| Salmonella Typhimurium str. LT2 | 432 | 575 | 110 | 154 | 7.27E-01 | 1.0518 | 0 |

**Table S11: Operons and the conditionally essential genes.** Each row corresponds to a given species or strain. Columns, from left to right, show species (strain) names, the number of metabolic genes that are neither conditionally essential nor belong to an operon $(--)$, the number of metabolic genes that are not conditionally essential but do belong to an operon $(-+)$, the number of metabolic genes that are conditionally essential, but do not belong to an operon $(+-)$, the number of metabolic genes that are both conditionally essential and belong to an operon $(++)$, the P value of a Fisher exact test on this data, the odds ratio (defined as the odds of being conditionally essential for operonic metabolic genes divided by the odds of being conditionally essential for non-operonic metabolic genes), and whether the null hypothesis of a lack of association between a gene's conditional essentiality and being on an operon is rejected (1) or not (0). In 46 of the 52 species (88.46%) the null hypothesis is rejected. Note that we consider a metabolic gene conditionally essential, if its deletion abolishes viability on at least one carbon source.

| | Essential genes belonging to an operon | | | | Essential genes not belonging to an operon | | | | All essential genes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Species(strains) | Hypothesis rejected | P-value | Kuiper test statistic | Critical value | Hypothesis rejected | P-value | Kuiper test statistic | Critical value | Hypothesis rejected | P-value | Kuiper test statistic | Critical value |
| *E. coli* K-12 MG1655 [iAF1260] | 1 | 1.22E-03 | 0.232 | 0.1777 | 0 | 1.38E-01 | 0.2121 | 0.2385 | 1 | 5.48E-03 | 0.1713 | 0.1435 |
| Methanosarcina barkeri str. Fusaro | 1 | 3.65E-02 | 0.2797 | 0.271 | 0 | 3.93E-01 | 0.1316 | 0.1769 | 0 | 2.53E-01 | 0.1211 | 0.1492 |
| Geobacter metallireducens GS-15 | 1 | 2.08E-06 | 0.22 | 0.1307 | 1 | 9.90E-04 | 0.248 | 0.1876 | 1 | 1.84E-05 | 0.1688 | 0.1078 |
| *E. coli* APEC O1 | 1 | 5.75E-07 | 0.303 | 0.1736 | 0 | 5.79E-01 | 0.1476 | 0.22 | 1 | 4.16E-05 | 0.2082 | 0.1371 |
| *E. coli* BL21(DE3) [iB21 1397] | 1 | 2.56E-02 | 0.165 | 0.1551 | 0 | 2.08E-01 | 0.2274 | 0.271 | 1 | 3.76E-03 | 0.1656 | 0.1354 |
| *E. coli* BW2952 | 1 | 3.63E-05 | 0.2183 | 0.143 | 0 | 5.78E-01 | 0.1442 | 0.2147 | 1 | 2.31E-04 | 0.1694 | 0.1197 |
| *E. coli* CFT073 | 1 | 4.38E-04 | 0.222 | 0.1617 | 0 | 2.76E-01 | 0.1529 | 0.191 | 1 | 5.52E-04 | 0.1687 | 0.1242 |
| *E. coli* O127:H6 | 1 | 2.06E-03 | 0.2113 | 0.1668 | 0 | 6.16E-02 | 0.2311 | 0.2362 | 1 | 2.15E-03 | 0.1733 | 0.1371 |
| *E. coli* 042 | 1 | 1.24E-03 | 0.2184 | 0.1676 | 1 | 4.61E-02 | 0.238 | 0.2362 | 1 | 7.55E-04 | 0.1839 | 0.1375 |
| *E. coli* 55989 | 1 | 2.65E-04 | 0.2324 | 0.1653 | 0 | 8.06E-02 | 0.231 | 0.2432 | 1 | 6.16E-04 | 0.1858 | 0.1375 |
| *E. coli* ABU 83972 | 1 | 1.42E-03 | 0.2168 | 0.1676 | 1 | 4.25E-02 | 0.2377 | 0.2339 | 1 | 1.17E-03 | 0.1792 | 0.1371 |
| *E. coli* B str. REL606 | 1 | 1.31E-03 | 0.2148 | 0.1653 | 0 | 1.27E-01 | 0.2186 | 0.2432 | 1 | 3.72E-03 | 0.1682 | 0.1375 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 1 | 2.19E-02 | 0.1742 | 0.1617 | 0 | 9.21E-02 | 0.2275 | 0.2432 | 1 | 6.43E-03 | 0.1599 | 0.1354 |
| *E. coli* BL21(DE3) [iECD1391] | 1 | 1.11E-02 | 0.1846 | 0.1624 | 0 | 1.07E-01 | 0.2214 | 0.241 | 1 | 2.63E-03 | 0.1692 | 0.1354 |
| *E. coli* DH1 [iEcDH1 1363] | 1 | 1.18E-03 | 0.216 | 0.1653 | 1 | 4.60E-02 | 0.2451 | 0.2432 | 1 | 6.30E-04 | 0.1856 | 0.1375 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 1 | 8.90E-03 | 0.1892 | 0.1638 | 0 | 2.47E-01 | 0.2041 | 0.2504 | 1 | 2.42E-02 | 0.1474 | 0.138 |
| *E. coli* E24377A | 1 | 3.74E-04 | 0.2298 | 0.1661 | 0 | 5.76E-02 | 0.2373 | 0.241 | 1 | 6.38E-04 | 0.1855 | 0.1375 |
| *E. coli* ED1a | 1 | 2.19E-03 | 0.2126 | 0.1684 | 0 | 1.42E-01 | 0.2036 | 0.2295 | 1 | 5.07E-03 | 0.164 | 0.1367 |
| *E. coli* O157:H7 | 1 | 1.08E-03 | 0.219 | 0.1668 | 1 | 4.48E-02 | 0.2409 | 0.2385 | 1 | 9.67E-04 | 0.1816 | 0.1375 |
| *E. coli* HS | 1 | 7.29E-04 | 0.2214 | 0.1653 | 0 | 1.06E-01 | 0.2237 | 0.2432 | 1 | 1.74E-03 | 0.1759 | 0.1375 |
| *E. coli* NA114 | 1 | 8.48E-06 | 0.2734 | 0.17 | 0 | 8.60E-02 | 0.2188 | 0.2317 | 1 | 2.84E-03 | 0.1716 | 0.138 |
| *E. coli* O103:H2 str. 12009 | 1 | 3.74E-04 | 0.2298 | 0.1661 | 0 | 1.79E-01 | 0.2065 | 0.241 | 1 | 1.37E-03 | 0.1782 | 0.1375 |
| *E. coli* O111:H- str. 11128 | 1 | 5.64E-04 | 0.2243 | 0.1653 | 0 | 5.81E-02 | 0.2394 | 0.2432 | 1 | 8.53E-04 | 0.1828 | 0.1375 |
| *E. coli* O26:H11 str. 11368 | 1 | 3.58E-04 | 0.2292 | 0.1653 | 0 | 1.00E-01 | 0.2252 | 0.2432 | 1 | 8.09E-04 | 0.1833 | 0.1375 |
| *E. coli* IHE3034 | 1 | 2.16E-03 | 0.2128 | 0.1684 | 0 | 7.99E-02 | 0.2206 | 0.2317 | 1 | 2.31E-03 | 0.1726 | 0.1371 |
| *E. coli* ATCC 8739 | 1 | 7.05E-04 | 0.2228 | 0.1661 | 0 | 8.84E-02 | 0.2264 | 0.241 | 1 | 1.15E-03 | 0.18 | 0.1375 |
| *E. coli* 536 | 1 | 1.90E-03 | 0.2143 | 0.1684 | 1 | 4.43E-02 | 0.2346 | 0.2317 | 1 | 1.31E-03 | 0.1782 | 0.1371 |
| *E. coli* O157:H7 str. Sakai | 1 | 1.09E-03 | 0.2189 | 0.1668 | 0 | 6.49E-02 | 0.232 | 0.2385 | 1 | 1.00E-03 | 0.1812 | 0.1375 |
| *E. coli* S88 | 1 | 1.34E-03 | 0.2175 | 0.1676 | 0 | 1.02E-01 | 0.2164 | 0.2339 | 1 | 2.07E-03 | 0.1737 | 0.1371 |
| *E. coli* SE11 | 1 | 2.85E-04 | 0.2316 | 0.1653 | 0 | 1.02E-01 | 0.2247 | 0.2432 | 1 | 6.58E-04 | 0.1852 | 0.1375 |
| *E. coli* SE15 | 1 | 4.79E-05 | 0.2168 | 0.1435 | 0 | 6.35E-01 | 0.1385 | 0.213 | 1 | 1.51E-03 | 0.1543 | 0.1197 |
| *E. coli* SMS-3-5 | 1 | 1.45E-04 | 0.2376 | 0.1646 | 0 | 7.84E-02 | 0.2341 | 0.2455 | 1 | 7.94E-04 | 0.1835 | 0.1375 |
| *E. coli* O157:H7 str. TW14359 | 1 | 1.82E-03 | 0.2138 | 0.1676 | 0 | 5.25E-02 | 0.235 | 0.2362 | 1 | 1.15E-03 | 0.1799 | 0.1375 |
| *E. coli* UMN026 | 1 | 4.11E-04 | 0.2277 | 0.1653 | 0 | 1.89E-01 | 0.2047 | 0.241 | 1 | 1.27E-03 | 0.1785 | 0.1371 |
| *E. coli* W [iECW_1372] | 1 | 2.31E-04 | 0.2328 | 0.1646 | 0 | 6.55E-02 | 0.2412 | 0.2479 | 1 | 2.09E-04 | 0.196 | 0.138 |
| *E. coli* KO11FL | 1 | 3.82E-04 | 0.2296 | 0.1661 | 0 | 5.48E-02 | 0.2408 | 0.2432 | 1 | 5.37E-04 | 0.1876 | 0.138 |
| *E. coli* ETEC H10407 | 1 | 3.24E-04 | 0.2303 | 0.1653 | 0 | 1.02E-01 | 0.2247 | 0.2432 | 1 | 7.23E-04 | 0.1843 | 0.1375 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *E. coli* O55:H7 str. CB9615 | 1 | 4.30E-04 | 0.2293 | 0.1668 | 1 | 4.48E-02 | 0.2409 | 0.2385 | 1 | 3.12E-04 | 0.192 | 0.1375 |
| *E. coli* K-12 MG1655 [iJO1366] | 1 | 3.03E-05 | 0.2198 | 0.143 | 0 | 6.38E-01 | 0.1404 | 0.2164 | 1 | 8.37E-04 | 0.1596 | 0.12 |
| *E. coli* K-12 MG1655 [iJR904] | 1 | 4.14E-03 | 0.237 | 0.1945 | 0 | 9.51E-01 | 0.113 | 0.2295 | 1 | 3.15E-02 | 0.1565 | 0.1498 |
| *E. coli* LF82 | 1 | 3.71E-09 | 0.332 | 0.1676 | 1 | 4.64E-02 | 0.2357 | 0.2339 | 1 | 2.06E-09 | 0.275 | 0.1371 |
| Mycobacterium tuberculosis H37Rv | 1 | 1.79E-02 | 0.1772 | 0.1617 | 0 | 8.87E-01 | 0.1007 | 0.1865 | 0 | 2.95E-01 | 0.097 | 0.123 |
| *E. coli* O83:H1 str. NRG 857C | 1 | 1.61E-03 | 0.2143 | 0.1668 | 1 | 4.73E-02 | 0.2375 | 0.2362 | 1 | 1.59E-03 | 0.1763 | 0.1371 |
| Shigella flexneri 2a str. 2457T | 1 | 3.69E-03 | 0.2014 | 0.1646 | 0 | 1.80E-01 | 0.1987 | 0.2317 | 1 | 1.13E-02 | 0.1533 | 0.135 |
| Shigella flexneri 5 str. 8401 | 1 | 3.16E-03 | 0.2042 | 0.1653 | 1 | 4.79E-02 | 0.2328 | 0.2317 | 1 | 9.77E-03 | 0.1554 | 0.1354 |
| *E. coli* UM146 | 1 | 1.20E-03 | 0.2187 | 0.1676 | 1 | 1.07E-02 | 0.269 | 0.2362 | 1 | 7.51E-04 | 0.184 | 0.1375 |
| *E. coli* UMNK88 | 1 | 1.21E-02 | 0.1867 | 0.1653 | 0 | 4.01E-01 | 0.1799 | 0.2432 | 0 | 1.73E-01 | 0.1185 | 0.1375 |
| *E. coli* UTI89 | 1 | 9.01E-04 | 0.2221 | 0.1676 | 0 | 7.54E-02 | 0.2241 | 0.2339 | 1 | 1.30E-03 | 0.1782 | 0.1371 |
| Klebsiella pneumoniae MGH78578 | 1 | 1.23E-02 | 0.2798 | 0.2479 | 0 | 7.29E-01 | 0.1751 | 0.2847 | 1 | 4.24E-02 | 0.1922 | 0.1887 |
| Bacillus subtilis str. 168 | 1 | 4.03E-04 | 0.2624 | 0.1898 | 0 | 5.51E-01 | 0.1895 | 0.2775 | 1 | 4.12E-04 | 0.218 | 0.1583 |
| *E. coli* O157:H7 str. EDL933 | 1 | 3.98E-04 | 0.2302 | 0.1668 | 1 | 4.33E-02 | 0.2418 | 0.2385 | 1 | 3.00E-04 | 0.1923 | 0.1375 |
| Salmonella Typhimurium str. LT2 | 1 | 2.37E-05 | 0.2988 | 0.1922 | 0 | 8.02E-02 | 0.2112 | 0.2218 | 1 | 3.02E-04 | 0.2048 | 0.1465 |

**Table S12: Operons may play an important role in the clustering of strictly essential genes.** In this analysis we divided the set of strictly non-essential genes into two groups: *i)* those belonging to an operon, and *ii)* those not belonging to an operon. Using Kuiper's test, we examined the clustering of *i)* the first group of strictly essential genes alone (columns 2-5; labeled as red in the first row), ii) the second group of strictly essential genes alone (columns 6-9; labeled as blue in the first row), and *iii)* all strictly essential genes together (columns 10-13; labeled as red in the first row). Each row corresponds to a bacterial species or strain. The first column is the species (strain) name, and in each of the three set of four columns, from left to right, columns show whether the null hypothesis of a uniform distribution of strictly essential genes is rejected by Kuiper's test (1) or not (0), the P-value of the test, Kuiper's test statistics, and the critical value of this statistic above which the null hypothesis is rejected. In 51 among the 55 species the null hypothesis is rejected, i.e., essential genes are significantly clustered. We consider a metabolic gene strictly essential, if its deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable. In 40 out of the 52 genomes (76.92%, colored blue) used in this analysis, the strictly essential genes acquired by horizontal gene transfer were significantly clustered (first group), but strictly essential genes not acquired by horizontal gene transfer (second group) were not significantly clustered. In the remaining 12 genomes (23.07%, colored black) strictly essential genes of both groups were significantly clustered.

| Species (strain) | Number of synthetic lethal gene pairs | Minimum genomic distance between synthetic lethal genes | Number of synthetic lethal pairs with genomic distance <50 |
|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 35 | 51 | 0 |
| Methanosarcina barkeri str. Fusaro | 71 | 0 | 14 |
| Geobacter metallireducens GS-15 | 89 | 0 | 18 |
| *E. coli* APEC O1 | 59 | 69 | 0 |
| *E. coli* BL21(DE3) [iB21 1397] | 40 | 51 | 0 |
| *E. coli* BW2952 | 68 | 65 | 0 |
| *E. coli* CFT073 | 191 | 0 | 3 |
| *E. coli* O127:H6 | 59 | 69 | 0 |
| *E. coli* 042 | 60 | 51 | 0 |
| *E. coli* 55989 | 60 | 51 | 0 |
| *E. coli* ABU 83972 | 59 | 69 | 0 |
| *E. coli* B str. REL606 | 60 | 51 | 0 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 40 | 50 | 0 |
| *E. coli* BL21(DE3) [iECD1391] | 40 | 51 | 0 |
| *E. coli* DH1 [iEcDH1 1363] | 60 | 53 | 0 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 61 | 54 | 0 |
| *E. coli* E24377A | 59 | 51 | 0 |
| *E. coli* ED1a | 59 | 64 | 0 |
| *E. coli* O157:H7 | 60 | 51 | 0 |
| *E. coli* HS | 60 | 51 | 0 |
| *E. coli* NA114 | 59 | 0 | 2 |
| *E. coli* O103:H2 str. 12009 | 61 | 0 | 3 |
| *E. coli* O111:H- str. 11128 | 60 | 52 | 0 |
| *E. coli* O26:H11 str. 11368 | 60 | 52 | 0 |
| *E. coli* IHE3034 | 59 | 67 | 0 |
| *E. coli* ATCC 8739 | 60 | 52 | 0 |
| *E. coli* 536 | 59 | 55 | 0 |
| *E. coli* O157:H7 str. Sakai | 60 | 51 | 0 |
| *E. coli* S88 | 59 | 68 | 0 |
| *E. coli* SE11 | 60 | 51 | 0 |
| *E. coli* SE15 | 67 | 67 | 0 |
| *E. coli* SMS-3-5 | 60 | 52 | 0 |
| *E. coli* O157:H7 str. TW14359 | 60 | 51 | 0 |
| *E. coli* UMN026 | 56 | 51 | 0 |
| *E. coli* W [iECW_1372] | 61 | 0 | 2 |
| *E. coli* KO11FL | 61 | 52 | 0 |
| *E. coli* ETEC H10407 | 60 | 51 | 0 |
| *E. coli* O55:H7 str. CB9615 | 60 | 51 | 0 |
| *E. coli* K-12 MG1655 [iJO1366] | 70 | 54 | 0 |
| *E. coli* K-12 MG1655 [iJR904] | 20 | 4 | 4 |
| *E. coli* LF82 | 59 | 0 | 19 |
| Mycobacterium tuberculosis H37Rv | 41 | 1 | 9 |
| *E. coli* O83:H1 str. NRG 857C | 59 | 68 | 0 |
| Shigella flexneri 2a str. 2457T | 62 | 63 | 0 |
| Staphylococcus aureus N315 | 21 | 0 | 5 |
| Shigella flexneri 5 str. 8401 | 60 | 67 | 0 |
| *E. coli* UM146 | 60 | 57 | 0 |
| *E. coli* UMNK88 | 60 | 0 | 2 |
| *E. coli* UTI89 | 59 | 57 | 0 |
| *E. coli* W [iWFL 1372] | 61 | 0 | 2 |
| *E. coli* str. K-12 W3110 | 60 | 54 | 0 |
| Klebsiella pneumoniae_MGH78578 | 37 | 0 | 9 |
| Bacillus subtilis str. 168 | 40 | 0 | 8 |
| *E. coli* O157:H7 str. EDL933 | 60 | 51 | 0 |
| Salmonella Typhimurium str. LT2 | 30 | 0 | 9 |

**Table S13: Strictly synthetic lethal gene pairs.** Each row corresponds to one of the 55 bacterial species or strains. Columns, from left to right, show the species name, the number of strictly synthetic lethal gene pairs, the distance between the synthetic lethal pairs with the shortest distance (smallest number of intervening genes) in the genome, and the number of strictly synthetic lethal gene pairs with distance below 50 intervening genes. In 41 of the genomes (82%), there are no strictly synthetic lethal gene pairs with distance below 50. We consider a pair of non-essential genes as strictly synthetic lethal if their simultaneous deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Species (strain) | (–) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 574538 | 48713 | 35 | 0 | 1.10E-01 | 0 | 0 |
| Methanosarcina barkeri str. Fusaro | 130875 | 21682 | 57 | 14 | 1.76E-01 | 1.4826 | 0 |
| Geobacter metallireducens GS-15 | 232023 | 26009 | 71 | 18 | 3.96E-03 | 2.2616 | 1 |
| *E. coli* APEC O1 | 614115 | 49954 | 59 | 0 | 2.20E-02 | 0 | 1 |
| *E. coli* BL21(DE3) [iB21 1397] | 636626 | 50712 | 40 | 0 | 1.18E-01 | 0 | 0 |
| *E. coli* BW2952 | 577882 | 46453 | 68 | 0 | 9.34E-03 | 0 | 1 |
| *E. coli* CFT073 | 561681 | 45881 | 188 | 3 | 4.91E-04 | 0.1954 | 1 |
| *E. coli* O127:H6 | 582591 | 48476 | 59 | 0 | 2.28E-02 | 0 | 1 |
| *E. coli* 042 | 616382 | 49993 | 60 | 0 | 2.26E-02 | 0 | 1 |
| *E. coli* 55989 | 634193 | 50782 | 60 | 0 | 2.21E-02 | 0 | 1 |
| *E. coli* ABU 83972 | 621948 | 50213 | 59 | 0 | 2.18E-02 | 0 | 1 |
| *E. coli* B str. REL606 | 633074 | 50731 | 60 | 0 | 2.21E-02 | 0 | 1 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 655898 | 51517 | 40 | 0 | 1.16E-01 | 0 | 0 |
| *E. coli* BL21(DE3) [iECD1391] | 632135 | 50521 | 40 | 0 | 1.18E-01 | 0 | 0 |
| *E. coli* DH1 [iEcDH1 1363] | 671776 | 52370 | 60 | 0 | 2.15E-02 | 0 | 1 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 763580 | 56199 | 61 | 0 | 2.13E-02 | 0 | 1 |
| *E. coli* E24377A | 648766 | 51511 | 59 | 0 | 2.14E-02 | 0 | 1 |
| *E. coli* ED1a | 576221 | 48123 | 59 | 0 | 1.38E-02 | 0 | 1 |
| *E. coli* O157:H7 | 560227 | 47466 | 60 | 0 | 1.39E-02 | 0 | 1 |
| *E. coli* HS | 624135 | 50346 | 60 | 0 | 2.24E-02 | 0 | 1 |
| *E. coli* NA114 | 603008 | 49586 | 57 | 2 | 3.23E-01 | 0.4267 | 0 |
| *E. coli* O103:H2 str. 12009 | 630872 | 50595 | 58 | 3 | 6.26E-01 | 0.645 | 0 |
| *E. coli* O111:H- str. 11128 | 631969 | 50667 | 60 | 0 | 2.22E-02 | 0 | 1 |
| *E. coli* O26:H11 str. 11368 | 662610 | 51940 | 60 | 0 | 2.16E-02 | 0 | 1 |
| *E. coli* IHE3034 | 604285 | 49452 | 59 | 0 | 2.21E-02 | 0 | 1 |
| *E. coli* ATCC 8739 | 677564 | 52612 | 60 | 0 | 2.14E-02 | 0 | 1 |
| *E. coli* 536 | 609776 | 49691 | 59 | 0 | 2.20E-02 | 0 | 1 |
| *E. coli* O157:H7 str. Sakai | 602108 | 49343 | 60 | 0 | 1.39E-02 | 0 | 1 |
| *E. coli* S88 | 605384 | 49497 | 59 | 0 | 2.21E-02 | 0 | 1 |
| *E. coli* SE11 | 654561 | 51645 | 60 | 0 | 2.17E-02 | 0 | 1 |
| *E. coli* SE15 | 576901 | 46318 | 67 | 0 | 9.27E-03 | 0 | 1 |
| *E. coli* SMS-3-5 | 653452 | 51566 | 60 | 0 | 2.17E-02 | 0 | 1 |
| *E. coli* O157:H7 str. TW14359 | 599925 | 49245 | 60 | 0 | 1.39E-02 | 0 | 1 |
| *E. coli* UMN026 | 637525 | 50970 | 56 | 0 | 2.10E-02 | 0 | 1 |
| *E. coli* W [iECW_1372] | 683337 | 52893 | 59 | 2 | 3.23E-01 | 0.4379 | 0 |
| *E. coli* KO11FL | 662537 | 52012 | 61 | 0 | 2.22E-02 | 0 | 1 |
| *E. coli* ETEC H10407 | 637566 | 50925 | 60 | 0 | 2.21E-02 | 0 | 1 |
| *E. coli* O55:H7 str. CB9615 | 582568 | 48498 | 60 | 0 | 1.38E-02 | 0 | 1 |
| *E. coli* K-12 MG1655 [iJO1366] | 623648 | 48502 | 70 | 0 | 9.41E-03 | 0 | 1 |
| *E. coli* K-12 MG1655 [iJR904] | 263757 | 32288 | 16 | 4 | 2.67E-01 | 2.0422 | 0 |
| *E. coli* LF82 | 579179 | 47402 | 40 | 19 | 3.58E-08 | 5.8038 | 1 |
| Mycobacterium tuberculosis H37Rv | 91476 | 15899 | 32 | 9 | 1.89E-01 | 1.6182 | 0 |
| *E. coli* O83:H1 str. NRG 857C | 611979 | 49787 | 59 | 0 | 2.20E-02 | 0 | 1 |
| Shigella flexneri 2a str. 2457T | 477361 | 43292 | 57 | 0 | 1.38E-2 | 0 | 1 |
| Staphylococcus aureus N315 | 130683 | 24699 | 16 | 5 | 3.64E-01 | 1.6534 | 0 |
| Shigella flexneri 5 str. 8401 | 476324 | 43306 | 60 | 0 | 8.96E-03 | 0 | 1 |
| *E. coli* UM146 | 621883 | 50277 | 60 | 0 | 2.24E-02 | 0 | 1 |
| *E. coli* UMNK88 | 660439 | 51722 | 58 | 2 | 3.23E-01 | 0.4403 | 0 |
| *E. coli* UTI89 | 610793 | 49823 | 59 | 0 | 2.20E-02 | 0 | 1 |
| *E. coli* W [iWFL 1372] | 683337 | 52893 | 59 | 2 | 3.23E-01 | 0.4379 | 0 |
| *E. coli* str. K-12 W3110 | 665915 | 52226 | 60 | 0 | 2.16E-02 | 0 | 1 |
| Klebsiella pneumoniae MGH78578 | 597269 | 51924 | 28 | 9 | 2.11E-03 | 3.6973 | 1 |
| Bacillus subtilis str. 168 | 231107 | 30579 | 32 | 8 | 1.32E-01 | 1.8894 | 0 |
| *E. coli* O157:H7 str. EDL933 | 609796 | 49670 | 60 | 0 | 2.27E-02 | 0 | 1 |
| Salmonella Typhimurium str. LT2 | 589406 | 49579 | 21 | 9 | 3.23E-04 | 5.095 | 1 |

**Table S14: Repulsion of the strictly synthetic lethal gene pairs.** Each row corresponds to one of the 55 bacterial species or strains. Columns, from left to right, show the species (strain) name, the number of non-essential metabolic gene pairs that are neither (strictly) synthetic lethal nor less than 50 metabolic genes apart ($--$), the number of non-essential metabolic gene pairs that are not (strictly) synthetic lethal but less than 50 metabolic genes apart ($+-$), the number of non-essential metabolic gene pairs that are (strictly) synthetic lethal and less than 50 metabolic genes apart ($-+$), the number of non-essential metabolic gene pairs that are both (strictly) synthetic lethal and less than 50 metabolic genes apart ($++$), the P value of Fisher's exact test, the odds ratio (the odds of being synthetic lethal for pairs of non-essential metabolic genes with distance below 50, divided by the odds of being synthetic lethal for pairs of non-essential metabolic genes with distance below 50), and whether (strictly) synthetic lethal gene pairs are in significant repulsion. We consider a pair of non-essential genes as strictly synthetic lethal if their simultaneous deletion abolishes viability on all carbon sources on which the wild-type metabolism is viable.

| Species (strain) | Number of synthetic lethal gene pairs | Minimum genomic distance between synthetic lethal genes | Number of synthetic lethal pairs with genomic distance <50 |
|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 696 | 0 | 46 |
| Methanosarcina barkeri str. Fusaro | 71 | 0 | 14 |
| Geobacter metallireducens GS-15 | 89 | 0 | 18 |
| *E. coli* APEC O1 | 640 | 0 | 39 |
| *E. coli* BL21(DE3) [iB21 1397] | 614 | 0 | 41 |
| *E. coli* BW2952 | 652 | 0 | 51 |
| *E. coli* CFT073 | 191 | 0 | 3 |
| *E. coli* O127:H6 | 720 | 0 | 52 |
| *E. coli* 042 | 684 | 0 | 45 |
| *E. coli* 55989 | 666 | 0 | 46 |
| *E. coli* ABU 83972 | 660 | 0 | 38 |
| *E. coli* B str. REL606 | 633 | 0 | 43 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 614 | 0 | 41 |
| *E. coli* BL21(DE3) [iECD1391] | 624 | 0 | 53 |
| *E. coli* DH1 [iEcDH1 1363] | 626 | 0 | 40 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 625 | 0 | 39 |
| *E. coli* E24377A | 640 | 0 | 55 |
| *E. coli* ED1a | 636 | 0 | 39 |
| *E. coli* O157:H7 | 699 | 0 | 45 |
| *E. coli* HS | 697 | 0 | 44 |
| *E. coli* NA114 | 647 | 0 | 34 |
| *E. coli* O103:H2 str. 12009 | 633 | 0 | 53 |
| *E. coli* O111:H- str. 11128 | 689 | 0 | 49 |
| *E. coli* O26:H11 str. 11368 | 658 | 0 | 48 |
| *E. coli* IHE3034 | 640 | 0 | 32 |
| *E. coli* ATCC 8739 | 620 | 0 | 52 |
| *E. coli* 536 | 658 | 0 | 40 |
| *E. coli* O157:H7 str. Sakai | 652 | 0 | 49 |
| *E. coli* S88 | 657 | 0 | 33 |
| *E. coli* SE11 | 642 | 0 | 49 |
| *E. coli* SE15 | 662 | 0 | 39 |
| *E. coli* SMS-3-5 | 650 | 0 | 47 |
| *E. coli* O157:H7 str. TW14359 | 651 | 0 | 47 |
| *E. coli* UMN026 | 686 | 0 | 47 |
| *E. coli* W [iECW_1372] | 644 | 0 | 49 |
| *E. coli* KO11FL | 652 | 0 | 51 |
| *E. coli* ETEC H10407 | 632 | 0 | 44 |
| *E. coli* O55:H7 str. CB9615 | 741 | 0 | 51 |
| *E. coli* K-12 MG1655 [iJO1366] | 642 | 0 | 47 |
| *E. coli* K-12 MG1655 [iJR904] | 348 | 0 | 23 |
| *E. coli* LF82 | 671 | 0 | 78 |
| Mycobacterium tuberculosis H37Rv | 41 | 1 | 9 |
| *E. coli* O83:H1 str. NRG 857C | 670 | 0 | 39 |
| Shigella flexneri 2a str. 2457T | 713 | 0 | 58 |
| Staphylococcus aureus N315 | 322 | 0 | 45 |
| Shigella flexneri 5 str. 8401 | 626 | 0 | 44 |
| *E. coli* UM146 | 633 | 0 | 32 |
| *E. coli* UMNK88 | 622 | 0 | 37 |
| *E. coli* UTI89 | 660 | 0 | 37 |
| *E. coli* W [iWFL 1372] | 644 | 0 | 49 |
| *E. coli* str. K-12 W3110 | 641 | 0 | 43 |
| Klebsiella pneumoniae_MGH78578 | 576 | 0 | 37 |
| Bacillus subtilis str. 168 | 552 | 0 | 51 |
| *E. coli* O157:H7 str. EDL933 | 653 | 0 | 50 |
| Salmonella Typhimurium str. LT2 | 226 | 0 | 42 |

**Table S15: Conditionally synthetic lethal gene pairs.** Each row corresponds to one of the 55 bacterial species or strains. Columns, from left to right, show the species (strain) name, the number of conditionally synthetic lethal gene pairs, the distance between the synthetic lethal pairs with the shortest distance (smallest number of intervening genes) in the genome, and the number of conditionally synthetic lethal gene pairs with distance below 50. In 54 genomes, conditionally synthetic lethal gene pairs exist whose member genes are adjacent in the genome. We consider a pair of non-essential genes as conditionally synthetic lethal if their simultaneous deletion abolishes viability on some but not all carbon sources.

| Species (strain) | (–) | (-+) | (+-) | (++) | P-value | Odds ratio | Hypothesis rejected |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 573923 | 48667 | 650 | 46 | 2.58E-01 | 0.8346 | 0 |
| Methanosarcina barkeri str. Fusaro | 130875 | 21682 | 57 | 14 | 1.76E-01 | 1.4826 | 0 |
| Geobacter metallireducens GS-15 | 232023 | 26009 | 71 | 18 | 3.96E-03 | 2.2616 | 1 |
| *E. coli* APEC O1 | 613572 | 49916 | 601 | 39 | 2.02E-01 | 0.7977 | 0 |
| *E. coli* BL21(DE3) [iB21 1397] | 636092 | 50672 | 573 | 41 | 5.88E-01 | 0.8982 | 0 |
| *E. coli* BW2952 | 577348 | 46403 | 601 | 51 | 7.09E-01 | 1.0558 | 0 |
| *E. coli* CFT073 | 561681 | 45881 | 188 | 3 | 4.91E-04 | 0.1954 | 1 |
| *E. coli* O127:H6 | 581981 | 48425 | 668 | 52 | 7.26E-01 | 0.9355 | 0 |
| *E. coli* 042 | 615802 | 49949 | 639 | 45 | 4.24E-01 | 0.8682 | 0 |
| *E. coli* 55989 | 633632 | 50737 | 620 | 46 | 7.11E-01 | 0.9266 | 0 |
| *E. coli* ABU 83972 | 621384 | 50176 | 622 | 38 | 1.03E-01 | 0.7566 | 0 |
| *E. coli* B str. REL606 | 632543 | 50689 | 590 | 43 | 5.96E-01 | 0.9095 | 0 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 655364 | 51477 | 573 | 41 | 6.41E-01 | 0.911 | 0 |
| *E. coli* BL21(DE3) [iECD1391] | 631603 | 50469 | 571 | 53 | 2.84E-01 | 1.1616 | 0 |
| *E. coli* DH1 [iEcDH1 1363] | 671249 | 52331 | 586 | 40 | 4.87E-01 | 0.8756 | 0 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 763054 | 56161 | 586 | 39 | 6.34E-01 | 0.9042 | 0 |
| *E. coli* E24377A | 648240 | 51456 | 585 | 55 | 2.25E-01 | 1.1844 | 0 |
| *E. coli* ED1a | 575682 | 48085 | 597 | 39 | 1.57E-01 | 0.7821 | 0 |
| *E. coli* O157:H7 | 559632 | 47422 | 654 | 45 | 2.04E-01 | 0.812 | 0 |
| *E. coli* HS | 623541 | 50303 | 653 | 44 | 2.79E-01 | 0.8352 | 0 |
| *E. coli* NA114 | 602452 | 49554 | 613 | 34 | 2.55E-02 | 0.6743 | 1 |
| *E. coli* O103:H2 str. 12009 | 630350 | 50545 | 580 | 53 | 3.62E-01 | 1.1396 | 0 |
| *E. coli* O111:H- str. 11128 | 631388 | 50619 | 640 | 49 | 8.27E-01 | 0.955 | 0 |
| *E. coli* O26:H11 str. 11368 | 662059 | 51893 | 610 | 48 | 9.40E-01 | 1.0039 | 0 |
| *E. coli* IHE3034 | 603735 | 49421 | 608 | 32 | 1.33E-02 | 0.643 | 1 |
| *E. coli* ATCC 8739 | 677055 | 52561 | 568 | 52 | 2.44E-01 | 1.1793 | 0 |
| *E. coli* 536 | 609216 | 49652 | 618 | 40 | 1.83E-01 | 0.7942 | 0 |
| *E. coli* O157:H7 str. Sakai | 601564 | 49295 | 603 | 49 | 1.00E+00 | 0.9916 | 0 |
| *E. coli* S88 | 604818 | 49465 | 624 | 33 | 1.19E-02 | 0.6466 | 1 |
| *E. coli* SE11 | 654027 | 51597 | 593 | 49 | 7.61E-01 | 1.0474 | 0 |
| *E. coli* SE15 | 576344 | 46280 | 623 | 39 | 1.38E-01 | 0.7796 | 0 |
| *E. coli* SMS-3-5 | 652908 | 51520 | 603 | 47 | 1.00E+00 | 0.9878 | 0 |
| *E. coli* O157:H7 str. TW14359 | 599380 | 49199 | 604 | 47 | 8.24E-01 | 0.948 | 0 |
| *E. coli* UMN026 | 636942 | 50923 | 639 | 47 | 6.61E-01 | 0.92 | 0 |
| *E. coli* W [iECW_1372] | 682801 | 52846 | 595 | 49 | 6.47E-01 | 1.064 | 0 |
| *E. coli* KO11FL | 661996 | 51962 | 601 | 51 | 5.97E-01 | 1.0811 | 0 |
| *E. coli* ETEC H10407 | 637037 | 50882 | 588 | 44 | 7.61E-01 | 0.9369 | 0 |
| *E. coli* O55:H7 str. CB9615 | 581937 | 48448 | 690 | 51 | 4.48E-01 | 0.8878 | 0 |
| *E. coli* K-12 MG1655 [iJO1366] | 623123 | 48455 | 595 | 47 | 8.79E-01 | 1.0158 | 0 |
| *E. coli* K-12 MG1655 [iJR904] | 263448 | 32269 | 325 | 23 | 9.52E-03 | 0.5778 | 1 |
| *E. coli* LF82 | 578626 | 47343 | 593 | 78 | 1.83E-04 | 1.6076 | 1 |
| Mycobacterium tuberculosis H37Rv | 91476 | 15899 | 32 | 9 | 1.89E-01 | 1.6182 | 0 |
| *E. coli* O83:H1 str. NRG 857C | 611406 | 49749 | 631 | 39 | 1.06E-01 | 0.7596 | 0 |
| Shigella flexneri 2a str. 2457T | 476763 | 43234 | 655 | 58 | 9.46E-01 | 0.9765 | 0 |
| Staphylococcus aureus N315 | 130422 | 24659 | 277 | 45 | 4.01E-01 | 0.8592 | 0 |
| Shigella flexneri 5 str. 8401 | 475801 | 43263 | 582 | 44 | 2.77E-01 | 0.8315 | 0 |
| *E. coli* UM146 | 621341 | 50246 | 601 | 32 | 1.89E-02 | 0.6584 | 1 |
| *E. coli* UMNK88 | 659912 | 51687 | 585 | 37 | 2.46E-01 | 0.8075 | 0 |
| *E. coli* UTI89 | 610228 | 49787 | 623 | 37 | 6.47E-02 | 0.7279 | 0 |
| *E. coli* W [iWFL 1372] | 682801 | 52846 | 595 | 49 | 6.47E-01 | 1.064 | 0 |
| *E. coli* str. K-12 W3110 | 665376 | 52184 | 598 | 43 | 6.48E-01 | 0.9168 | 0 |
| Klebsiella pneumoniae MGH78578 | 596758 | 51896 | 539 | 37 | 1.91E-01 | 0.7894 | 0 |
| Bacillus subtilis str. 168 | 230638 | 30536 | 501 | 51 | 7.35E-02 | 0.7689 | 0 |
| *E. coli* O157:H7 str. EDL933 | 609252 | 49621 | 603 | 50 | 8.82E-01 | 1.0181 | 0 |
| Salmonella Typhimurium str. LT2 | 589243 | 49546 | 184 | 42 | 1.28E-07 | 2.7147 | 1 |

**Table S16: Repulsion of conditionally synthetic lethal gene pairs.** Each row corresponds to one of the 55 bacterial species or strains. Columns, from left to right, show the species (strain) name, the number of non-essential metabolic gene pairs that are neither (conditionally) synthetic lethal nor less than 50 metabolic genes apart $(--)$, the number of non-essential metabolic gene pairs that are not (conditionally) synthetic lethal but less than 50 metabolic genes apart $(+-)$, the number of non-essential metabolic gene pairs that are (conditionally) synthetic lethal and less than 50 metabolic genes apart $(-+)$, the number of non-essential metabolic gene pairs that are both (conditionally) synthetic lethal and less than 50 metabolic genes apart $(++)$, the P value of Fisher's exact test, the odds ratio (the odds of being (conditionally) synthetic lethal among the pairs of non-essential metabolic genes with less than 50 metabolic genes apart divided by the odds of being (conditionally) synthetic lethal among the pairs of non-essential metabolic genes with more than or equal to 50 metabolic genes apart), and whether (conditionally) synthetic lethal gene pairs are in significant repulsion. We consider a pair of non-essential genes as conditionally synthetic lethal if their simultaneous deletion abolishes viability on some but not all carbon sources.

| Species (strain) | Number of clusters | Average length | Max length | No. Strictly non-essential | Fraction strictly non-essential | No. Conditionally non-essential | Fraction conditionally non-essential |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 148 | 6.38 | 48 | 147 | 0.99 | 148 | 1 |
| Methanosarcina barkeri str. Fusaro | 81 | 6.89 | 33 | 78 | 0.96 | 78 | 0.96 |
| Geobacter metallireducens GS-15 | 103 | 7.09 | 92 | 98 | 0.95 | 98 | 0.95 |
| *E. coli* APEC O1 | 166 | 6.04 | 47 | 163 | 0.98 | 165 | 0.99 |
| *E. coli* BL21(DE3) [iB21 1397] | 154 | 6.62 | 47 | 151 | 0.98 | 153 | 0.99 |
| *E. coli* BW2952 | 170 | 5.72 | 30 | 167 | 0.98 | 169 | 0.99 |
| *E. coli* CFT073 | 107 | 10.4 | 70 | 104 | 0.97 | 104 | 0.97 |
| *E. coli* O127:H6 | 153 | 6.41 | 45 | 150 | 0.98 | 152 | 0.99 |
| *E. coli* 042 | 153 | 6.51 | 45 | 150 | 0.98 | 152 | 0.99 |
| *E. coli* 55989 | 157 | 6.45 | 66 | 155 | 0.99 | 156 | 0.99 |
| *E. coli* ABU 83972 | 155 | 6.52 | 47 | 152 | 0.98 | 154 | 0.99 |
| *E. coli* B str. REL606 | 154 | 6.57 | 47 | 151 | 0.98 | 153 | 0.99 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 154 | 6.73 | 47 | 150 | 0.97 | 152 | 0.99 |
| *E. coli* BL21(DE3) [iECD1391] | 155 | 6.55 | 47 | 152 | 0.98 | 154 | 0.99 |
| *E. coli* DH1 [iEcDH1 1363] | 151 | 6.95 | 47 | 147 | 0.97 | 149 | 0.99 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 150 | 7.52 | 97 | 147 | 0.98 | 149 | 0.99 |
| *E. coli* E24377A | 154 | 6.68 | 66 | 150 | 0.97 | 154 | 1 |
| *E. coli* ED1a | 156 | 6.21 | 47 | 153 | 0.98 | 155 | 0.99 |
| *E. coli* O157:H7 | 152 | 6.24 | 44 | 151 | 0.99 | 151 | 0.99 |
| *E. coli* HS | 154 | 6.51 | 47 | 152 | 0.99 | 153 | 0.99 |
| *E. coli* NA114 | 155 | 6.39 | 47 | 152 | 0.98 | 154 | 0.99 |
| *E. coli* O103:H2 str. 12009 | 154 | 6.62 | 66 | 152 | 0.99 | 153 | 0.99 |
| *E. coli* O111:H- str. 11128 | 154 | 6.62 | 65 | 150 | 0.97 | 153 | 0.99 |
| *E. coli* O26:H11 str. 11368 | 156 | 6.67 | 66 | 154 | 0.99 | 155 | 0.99 |
| *E. coli* IHE3034 | 157 | 6.33 | 47 | 154 | 0.98 | 156 | 0.99 |
| *E. coli* ATCC 8739 | 151 | 6.98 | 47 | 147 | 0.97 | 149 | 0.99 |
| *E. coli* 536 | 158 | 6.32 | 47 | 155 | 0.98 | 157 | 0.99 |
| *E. coli* O157:H7 str. Sakai | 154 | 6.45 | 66 | 152 | 0.99 | 153 | 0.99 |
| *E. coli* S88 | 154 | 6.47 | 46 | 151 | 0.98 | 153 | 0.99 |
| *E. coli* SE11 | 153 | 6.76 | 67 | 150 | 0.98 | 152 | 0.99 |
| *E. coli* SE15 | 178 | 5.4 | 36 | 175 | 0.98 | 177 | 0.99 |
| *E. coli* SMS-3-5 | 154 | 6.68 | 47 | 151 | 0.98 | 153 | 0.99 |
| *E. coli* O157:H7 str. TW14359 | 154 | 6.42 | 66 | 153 | 0.99 | 153 | 0.99 |
| *E. coli* UMN026 | 156 | 6.47 | 47 | 153 | 0.98 | 156 | 1 |
| *E. coli* W [iECW 1372] | 152 | 6.98 | 62 | 150 | 0.99 | 151 | 0.99 |
| *E. coli* KO11FL | 153 | 6.81 | 66 | 150 | 0.98 | 151 | 0.99 |
| *E. coli* ETEC H10407 | 152 | 6.73 | 47 | 149 | 0.98 | 151 | 0.99 |
| *E. coli* O55:H7 str. CB9615 | 154 | 6.24 | 44 | 152 | 0.99 | 153 | 0.99 |
| *E. coli* K-12 MG1655 [iJO1366] | 175 | 5.69 | 32 | 174 | 0.99 | 175 | 1 |
| *E. coli* K-12 MG1655 [iJR904] | 135 | 4.39 | 29 | 134 | 0.99 | 135 | 1 |
| *E. coli* LF82 | 129 | 7.71 | 60 | 111 | 0.86 | 121 | 0.94 |
| Mycobacterium tuberculosis H37Rv | 104 | 4.46 | 22 | 104 | 1 | 104 | 1 |
| *E. coli* O83:H1 str. NRG 857C | 152 | 6.61 | 47 | 149 | 0.98 | 151 | 0.99 |
| Shigella flexneri 2a str. 2457T | 148 | 6.04 | 29 | 145 | 0.98 | 146 | 0.99 |
| Staphylococcus aureus N315 | 82 | 5.59 | 21 | 80 | 0.98 | 81 | 0.99 |
| Shigella flexneri 5 str. 8401 | 149 | 6.03 | 40 | 147 | 0.99 | 148 | 0.99 |
| *E. coli* UM146 | 155 | 6.52 | 47 | 152 | 0.98 | 154 | 0.99 |
| *E. coli* UMNK88 | 168 | 6.21 | 47 | 165 | 0.98 | 167 | 0.99 |
| *E. coli* UTI89 | 157 | 6.37 | 47 | 154 | 0.98 | 156 | 0.99 |
| *E. coli* W [iWFL 1372] | 152 | 6.98 | 62 | 150 | 0.99 | 151 | 0.99 |
| *E. coli* str. K-12 W3110 | 155 | 6.73 | 47 | 153 | 0.99 | 154 | 0.99 |
| Klebsiella pneumoniae MGH78578 | 121 | 8.1 | 39 | 116 | 0.96 | 118 | 0.98 |
| Bacillus subtilis str. 168 | 108 | 5.38 | 34 | 104 | 0.96 | 107 | 0.99 |
| *E. coli* O157:H7 str. EDL933 | 151 | 6.62 | 67 | 149 | 0.99 | 150 | 0.99 |
| Salmonella Typhimurium str. LT2 | 139 | 7.24 | 61 | 135 | 0.97 | 138 | 0.99 |

**Table S17: Non-essential clusters of strictly non-essential genes.** Each row corresponds to one of the 55 bacterial species or strains. Columns, show the species (strain) name (first column), the number of the clusters of strictly non-essential genes (second column), the average length of the clusters of strictly non-essential genes (third column), the length of the largest cluster of non-essential genes (fourth column), the number (fifth column) and fraction (sixth column) of strictly non-essential clusters of strictly non-essential genes, and the number (seventh column) and fraction (eighth column) of conditionally non-essential clusters of strictly non-essential genes. We consider a metabolic gene as conditionally non-essential if its deletion does not abolish viability on at least one carbon source, and consider a metabolic gene as strictly non-essential if its deletion does not abolish viability on any carbon source. A *cluster of conditionally non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that lies between two nearest non-adjacent strictly essential metabolic genes. Likewise, a *cluster of strictly non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that lies between two nearest non-adjacent conditionally essential metabolic genes. Finally, we consider a cluster of strictly non-essential genes as a *strictly non-essential cluster of strictly non-essential genes* if simultaneous deletion of all the genes in the cluster does not abolish viability on any carbon source, and we consider a cluster of strictly non-essential genes as a*conditionally non-essential cluster of strictly non-essential genes* if simultaneous deletion of all the genes in the cluster does not abolish viability on at least one carbon source.

| Species (strain) | Number of clusters | Average length | Max length | No. Strictly non-essential | Fraction strictly non-essential | No. Conditionally non-essential | Fraction conditionally non-essential |
|---|---|---|---|---|---|---|---|
| *E. coli* K-12 MG1655 [iAF1260] | 81 | 13.79 | 91 | 39 | 0.48 | 76 | 0.94 |
| Methanosarcina barkeri str. Fusaro | 81 | 6.89 | 33 | 78 | 0.96 | 78 | 0.96 |
| Geobacter metallireducens GS-15 | 103 | 7.09 | 92 | 98 | 0.95 | 98 | 0.95 |
| *E. coli* APEC O1 | 94 | 12.28 | 73 | 48 | 0.51 | 88 | 0.94 |
| *E. coli* BL21(DE3) [iB21 1397] | 87 | 13.49 | 74 | 47 | 0.54 | 81 | 0.93 |
| *E. coli* BW2952 | 109 | 10.28 | 63 | 67 | 0.61 | 103 | 0.94 |
| *E. coli* CFT073 | 107 | 10.4 | 70 | 104 | 0.97 | 104 | 0.97 |
| *E. coli* O127:H6 | 86 | 13.08 | 76 | 47 | 0.55 | 78 | 0.91 |
| *E. coli* 042 | 86 | 13.44 | 71 | 46 | 0.53 | 79 | 0.92 |
| *E. coli* 55989 | 86 | 13.63 | 70 | 46 | 0.53 | 79 | 0.92 |
| *E. coli* ABU 83972 | 86 | 13.5 | 72 | 46 | 0.53 | 78 | 0.91 |
| *E. coli* B str. REL606 | 86 | 13.62 | 71 | 45 | 0.52 | 79 | 0.92 |
| *E. coli* BL21-Gold(DE3)pLysS AG | 87 | 13.69 | 73 | 47 | 0.54 | 81 | 0.93 |
| *E. coli* BL21(DE3) [iECD1391] | 87 | 13.45 | 73 | 47 | 0.54 | 81 | 0.93 |
| *E. coli* DH1 [iEcDH1 1363] | 85 | 14.18 | 80 | 44 | 0.52 | 79 | 0.93 |
| *E. coli* DH1 [iECDH1ME8569 1439] | 84 | 15.27 | 98 | 43 | 0.51 | 78 | 0.93 |
| *E. coli* E24377A | 86 | 13.76 | 74 | 46 | 0.53 | 80 | 0.93 |
| *E. coli* ED1a | 87 | 12.86 | 70 | 47 | 0.54 | 80 | 0.92 |
| *E. coli* O157:H7 | 86 | 12.84 | 67 | 45 | 0.52 | 80 | 0.93 |
| *E. coli* HS | 86 | 13.52 | 74 | 46 | 0.53 | 79 | 0.92 |
| *E. coli* NA114 | 85 | 13.46 | 86 | 48 | 0.56 | 77 | 0.91 |
| *E. coli* O103:H2 str. 12009 | 86 | 13.59 | 66 | 48 | 0.56 | 79 | 0.92 |
| *E. coli* O111:H- str. 11128 | 86 | 13.6 | 91 | 48 | 0.56 | 79 | 0.92 |
| *E. coli* O26:H11 str. 11368 | 86 | 13.92 | 66 | 46 | 0.53 | 80 | 0.93 |
| *E. coli* IHE3034 | 86 | 13.31 | 73 | 46 | 0.53 | 78 | 0.91 |
| *E. coli* ATCC 8739 | 85 | 14.24 | 75 | 44 | 0.52 | 79 | 0.93 |
| *E. coli* 536 | 87 | 13.22 | 72 | 47 | 0.54 | 79 | 0.91 |
| *E. coli* O157:H7 str. Sakai | 87 | 13.14 | 66 | 46 | 0.53 | 81 | 0.93 |
| *E. coli* S88 | 86 | 13.33 | 72 | 46 | 0.53 | 78 | 0.91 |
| *E. coli* SE11 | 85 | 14 | 70 | 45 | 0.53 | 78 | 0.92 |
| *E. coli* SE15 | 109 | 10.26 | 63 | 67 | 0.61 | 102 | 0.94 |
| *E. coli* SMS-3-5 | 86 | 13.83 | 74 | 45 | 0.52 | 79 | 0.92 |
| *E. coli* O157:H7 str. TW14359 | 87 | 13.11 | 66 | 46 | 0.53 | 81 | 0.93 |
| *E. coli* UMN026 | 87 | 13.48 | 71 | 45 | 0.52 | 82 | 0.94 |
| *E. coli* W [iECW 1372] | 85 | 14.29 | 73 | 46 | 0.54 | 79 | 0.93 |
| *E. coli* KO11FL | 86 | 13.92 | 74 | 46 | 0.53 | 79 | 0.92 |
| *E. coli* ETEC H10407 | 86 | 13.66 | 72 | 46 | 0.53 | 79 | 0.92 |
| *E. coli* O55:H7 str. CB9615 | 86 | 13.08 | 67 | 44 | 0.51 | 79 | 0.92 |
| *E. coli* K-12 MG1655 [iJO1366] | 110 | 10.55 | 65 | 65 | 0.59 | 106 | 0.96 |
| *E. coli* K-12 MG1655 [iJR904] | 71 | 10.86 | 52 | 28 | 0.39 | 68 | 0.96 |
| *E. coli* LF82 | 70 | 16.33 | 86 | 33 | 0.47 | 57 | 0.81 |
| Mycobacterium tuberculosis H37Rv | 104 | 4.46 | 22 | 104 | 1 | 104 | 1 |
| *E. coli* O83:H1 str. NRG 857C | 86 | 13.4 | 72 | 46 | 0.53 | 78 | 0.91 |
| Shigella flexneri 2a str. 2457T | 89 | 11.51 | 70 | 50 | 0.56 | 83 | 0.93 |
| Staphylococcus aureus N315 | 36 | 15.58 | 64 | 15 | 0.42 | 29 | 0.81 |
| Shigella flexneri 5 str. 8401 | 88 | 11.6 | 68 | 51 | 0.58 | 82 | 0.93 |
| *E. coli* UM146 | 86 | 13.5 | 71 | 47 | 0.55 | 78 | 0.91 |
| *E. coli* UMNK88 | 105 | 11.38 | 76 | 64 | 0.61 | 98 | 0.93 |
| *E. coli* UTI89 | 87 | 13.23 | 74 | 47 | 0.54 | 79 | 0.91 |
| *E. coli* W [iWFL 1372] | 85 | 14.29 | 73 | 46 | 0.54 | 79 | 0.93 |
| *E. coli* str. K-12 W3110 | 86 | 13.95 | 77 | 45 | 0.52 | 79 | 0.92 |
| Klebsiella pneumoniae MGH78578 | 54 | 21.24 | 127 | 21 | 0.39 | 48 | 0.89 |
| Bacillus subtilis str. 168 | 52 | 13.94 | 72 | 25 | 0.48 | 46 | 0.88 |
| *E. coli* O157:H7 str. EDL933 | 86 | 13.37 | 67 | 46 | 0.53 | 80 | 0.93 |
| Salmonella Typhimurium str. LT2 | 77 | 14.71 | 99 | 41 | 0.53 | 71 | 0.92 |

**Table S18: Non-essential clusters of conditionally non-essential genes.** Each row corresponds to one of the 55 bacterial species or strains. Columns, show the species (strain) name (first column), the number of the clusters of conditionally non-essential genes (second column), the average length of the clusters of conditionally non-essential genes (third column), the length of the largest cluster of non-essential genes (fourth column), the number (fifth column) and fraction (sixth column) of strictly non-essential clusters of conditionally non-essential genes, and the number (seventh column) and fraction (eighth column) of conditionally non-essential clusters of conditionally non-essential genes. We consider a metabolic gene as conditionally non-essential if its deletion does not abolish viability on at least one carbon source, and consider a metabolic gene as strictly non-essential if its deletion does not abolish viability on any carbon source. A *cluster of conditionally non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that lies between two nearest non-adjacent strictly essential metabolic genes. Likewise, a *cluster of strictly non-essential metabolic genes* is a set of consecutive non-essential metabolic genes that lies between two nearest non-adjacent conditionally essential metabolic genes. Finally, we consider a cluster of conditionally non-essential genes as *strictly non-essential cluster of conditionally non-essential genes* if simultaneous deletion of all the genes in the cluster does not abolish viability on any carbon sources, and we consider a cluster of conditionally non-essential genes as *conditionally non-essential cluster of conditionally non-essential genes* if simultaneous deletion of all the genes in the cluster does not abolish viability on at least one carbon source.

# Chapter 4:

# Constraint and contingency pervade the emergence of novel phenotypes in complex metabolic systems

Sayed-Rzgar Hosseini and Andreas Wagner

The content of this chapter has been published as:
Hosseini, S.-R., and A. Wagner. 2017. Constraint and Contingency Pervade the Emergence of Novel Phenotypes in Complex Metabolic Systems. Biophys. J. 113: 690–701.
DOI: 10.1016/j.bpj.2017.06.034

## 4.1. Abstract

An evolutionary constraint is a bias or limitation in phenotypic variation that a biological system produces. We know examples of such constraints, but we have no systematic understanding about their extent and causes for any one biological system. We here study metabolisms, genomically encoded complex networks of enzyme-catalyzed biochemical reactions, and the constraints they experience in bringing forth novel phenotypes that allow survival on novel carbon sources. Our computational approach does not limit us to analyzing constrained variation in any one organism, but allows us to quantify constraints experienced by any metabolism. Specifically, we study metabolisms that are viable on one of 50 different carbon sources, and quantify how readily alterations of their chemical reactions create the ability to survive on a novel carbon source. We find that some metabolic phenotypes are much less likely to originate than others. For example, metabolisms viable on D-glucose are 1,835 times more likely to give rise to metabolisms viable on D-fructose than on acetate. Likewise, we observe that some novel metabolic phenotypes are more contingent on parental phenotypes than others. Biochemical similarities among carbon sources can help explain the causes of these constraints. In addition, we study metabolisms that can be produced by recombination among 55 metabolisms of different bacterial strains or species, and show that their novel phenotypes are also contingent on and constrained by "parental" genotypes. Our analysis is the first to systematically quantify the incidence of constrained evolution in a broad class of biological system that is central to life and its evolution.

## 4.2. Introduction

Individual organisms or populations cannot produce every conceivable kind of phenotypic variation. In other words, phenotypic evolution is to some extent constrained. More precisely, an evolutionary constraint is a bias or limitation in the emergence of phenotypic variation in a given biological system (1). Examples of constraints on the organismal level include the absence of photosynthesis in higher animals, the absence of birds that can give birth to live young instead of to eggs, the general lack of teeth in the lower jaw of frogs, and the absence of palm trees in cold climates (1, 2). Other examples include constrained variation in segment number, orientation and identity in the fruit fly *Drosophila melanogaster* (3), and correlations among different characters, such as in allometric scaling (4). Molecular examples of phenotypic constraints include the absence of L-isomers in the 20 amino acids found in natural proteins (5), and a limited number of possible protein folds caused by the packing requirements of hydrophobic amino-acids (6). It is useful to distinguish between *absolute* constraints, which occur when some phenotype cannot be produced, and *relative* constraints, when some phenotypes are more likely to arise than others.

A closely related concept is that of contingency. We speak of contingency when the origin of a novel phenotype depends on the history of a population, and specifically on pre-existing genotypes or phenotypes (7, 8). For example, experimental evolution of *Escherichia coli* has shown that the emergence of citrate-utilization as a novel metabolic phenotype is contingent on the genetic history of a population (9). Analogously to constraints, one can distinguish between phenotypes that are absolutely or relatively contingent on evolutionary history. Although many anecdotal examples of constraints and contingent evolution exist, such examples do not allow one to quantify the potential for either phenomenon in any one class of biological system. We here undertake such a quantification using a computational approach applied to metabolic systems, which are ideal for this purpose for several reasons.

First, metabolic systems, and especially those of microbes, are an abundant source of new adaptations and innovations – qualitatively new adaptations. Especially important innovations are those that allow an organism to extract energy and chemical elements from new molecules, which can help it survive in new habitats. For instance, microorganisms have acquired the ability to utilize many non-natural

substances, such as polychlorinated biphenyls, chlorobenzenes, organic solvents, synthetic pesticides, and even antibiotics as food (10–14).

Second, experimentally-validated computational methods such as flux balance analysis (FBA) provide efficient means to systematically predict metabolic phenotypes – the ability of an organism to survive on specific nutrients – from information about metabolic genotypes (15, 16). A metabolic genotype is the part of a genome encoding metabolic enzymes. However, computational analyses of metabolic systems, often use a more abstract and compact representation of such a genotype, referring to it as the collection of chemical reactions that a metabolic reaction network is able to catalyze (17–26).

Third, in metabolic systems, we are not restricted to studying the metabolism of any one organism, together with the constraints and contingencies it may be subject to. Instead, we can study the *potential* for contingency and constraint in entire classes of metabolic systems. To do so, we take advantage of Markov Chain Monte Carlo algorithms ((21, 23) (see methods)) that allow us to create large numbers of metabolisms. Each such metabolism is a complex network of chemical reactions with a given phenotype, but its complement of metabolic reactions is otherwise sampled at random from a "universe" of metabolic reactions that are known to exist among prokaryotes (see Methods). We refer to such metabolisms as random viable metabolic networks. The phenotypes we study are viability phenotypes, and specifically a metabolism's ability to synthesize all essential biomass precursors in a minimal medium that harbors only a single carbon source. We consider 50 such carbon sources, i.e., 50 different metabolic phenotypes.

When analyzing phenotypic variability, it is important to consider the kinds of genotype changes that cause this variability. We focus on recombination-like processes as a means for genotypic change, and do so for two reasons. First, recombination is a ubiquitous force of genetic change, not only in eukaryotes but also in prokaryotes whose genomes are being continually reorganized through horizontal gene transfer. Second, in contrast to smaller scale genetic change, such as point mutations, recombination causes larger-scale genetic change with greater potential to create novel phenotypes (27–32). Thus, if we found that phenotypic evolution was constrained when recombination causes genotypic change, it would be even more constrained if point mutations caused such change.

In our simulations, we generated one thousand parental pairs of random viable metabolic networks for each of the 50 carbon utilization phenotypes. For each one of these 50,000 parental pairs, using a recombination-like process that mimics horizontal gene transfer in bacteria (see methods), we generated 1,000 "offspring" to obtain 50 million recombinant metabolic networks. We focused on those recombinants that did not only retain viability on their parental carbon source, but also gained viability on at least one novel carbon source. For brevity, we will also refer to them as "innovative offspring". We analyze their phenotypes and how they depend on parental phenotypes. In addition, we also study recombination among metabolic networks of 55 bacterial species or strains.

We find little evidence for absolute constraints and contingencies. That is, the metabolic phenotypes we consider can be brought forth through recombination among some parental metabolic networks. However, relatively constraints and contingencies are pervasive. Differences in the biochemical relatedness of carbon sources, and the ensuing correlations among different carbon usage phenotypes can help explain some of these constraints and contingencies.

## 4.3. Results

**All metabolic phenotypes can emerge through recombination.**

Our first analysis focused on the perhaps most fundamental question regarding absolute constraints: Do some parental phenotypes not give rise to any offspring with novel phenotypes? To find out, we quantified for each carbon source $C_i$ and for each of the 1,000 parental pairs viable on $C_i$, the number $N_{C_i\rightarrow}$ of offspring gaining viability on *some* new carbon source ($C_j, j \neq i$), among their $10^6$ recombinant offspring (with $n$ = 10 altered reactions relative to the parents). Fig. 1a shows the distribution of this number, demonstrating that offspring with metabolic innovations can emerge from each of the 50 carbon usage phenotypes we analyzed. However, we also note that the number of offspring with a metabolic innovation varies greatly among different carbon usage phenotypes, ranging from 1,433 for parents viable on adenosine to 61,835 for parents viable on D-galactose (per million offspring). We repeated this analysis by varying the number of reactions ($n$) altered during recombination, which shows that the relative abundance and the ranking of carbon sources in terms of the frequency of innovation stays almost the same for various $n$

((Figs. S1 and S2); $n$ = 10 and 20, Spearman's $R$ = 0.9982; $P < 10^{-60}$; $n$ = 10 and 30, Spearman's $R$ = 0.9750; $P < 10^{-33}$.) In sum, all parental phenotypes we consider can give rise to metabolic innovations.

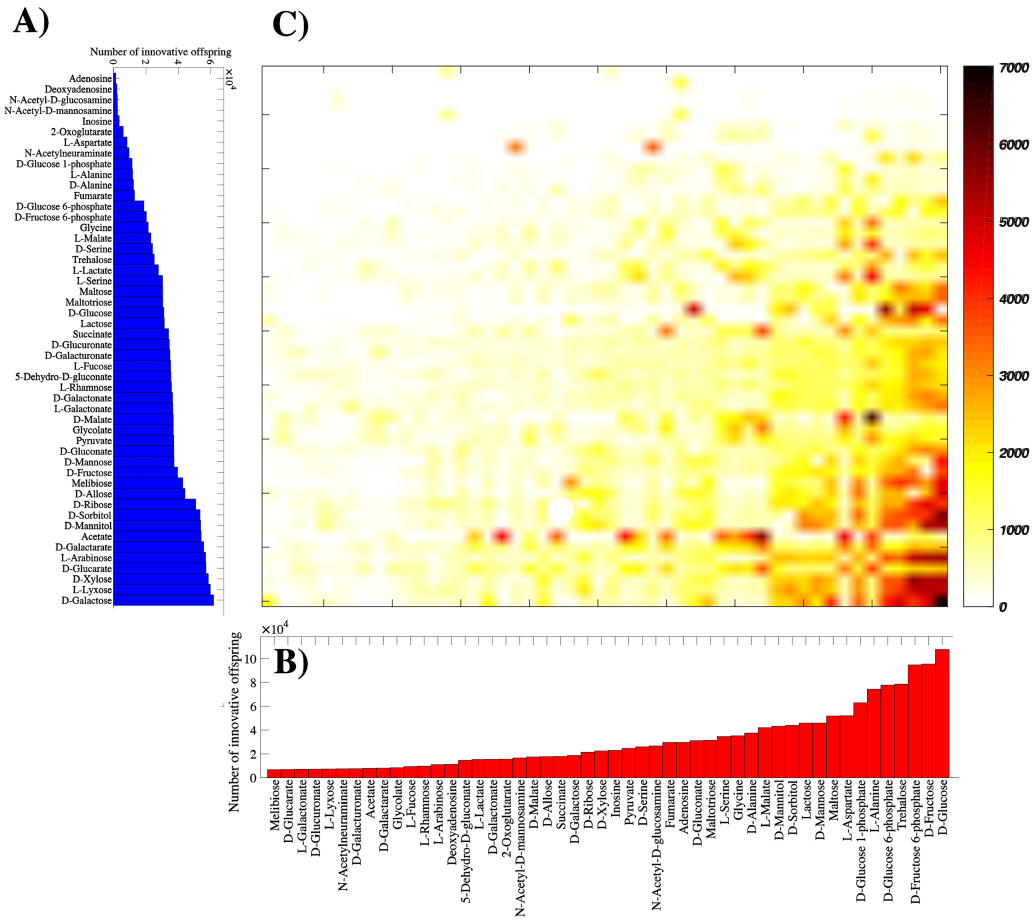



**Figure 1: Recombination can create all 50 carbon-use phenotypes considered here. A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the horizontal axis. **C)** Number of innovative recombinant (per million offspring, coded according to the color legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks during recombination.

Next, we asked whether different carbon usage phenotypes differ in their propensity to be found as novel offspring phenotypes, regardless of the parental phenotype. Fig. 1b shows that this is indeed the case. But while all 50 carbon-usage phenotypes appear in the innovative offspring we analyzed, their prevalence ($N_{\rightarrow C_i}$) varies by a factor 16 among carbon sources, ranging from 6,783 innovative offspring gaining viability on melibiose to 107,784 gaining viability on D-glucose (among $50\times10^6$ recombinant offspring, and a total of 1,556,237 innovative offspring). This variability is similarly great with a number ($n$) of recombined reactions different from $n$ = 10 (Figs. S1 and S2).We noted a negative correlation between $N_{C_i\rightarrow}$ and $N_{\rightarrow C_i}$ (Fig. S3), i.e., carbon-usage phenotypes that give rise to more innovative offspring are found less frequently as products of recombinational innovation.

Finally, Fig. 1c shows the variability among different pairs of carbon sources in terms of their propensity for generating innovative offspring. In 2,038 pairs (81.52% among the possible 2,500 pairs of carbon sources ($C_i, C_j$), fewer than 1,000 innovative recombinant (among one million offspring) gain viability on $C_j$ from recombination between parents viable on $C_i$, and only in 17 pairs (0.68%) more than 5,000 innovative offspring emerge. The largest number of innovative offspring (7,071) emerges when parents viable exclusively on D-galactose give rise to offspring that gain viability on D-glucose.

To find out whether parental genotypic distance and the number of reactions in a metabolic network might affect our observations, we repeated our analyses with more divergent parents ($D$ = 1,000) and smaller metabolic networks (1,800 and 1,600 reactions, as opposed to the 2,079 reactions identical to the number in *E. coli*, which we had used so far). Although recombination gives rise to fewer innovative offspring at higher $D$ and for smaller networks (Fig. S4), the general patterns (Figs. S5, S6, and S7) remain similar to that of Fig. 1.

Also, we had so far recombined parents that were viable on the same carbon source. To find out whether this could affect our observations, we generated recombinational offspring where one parent is viable on glucose, and the other is viable on a different carbon source. We found that recombination again results in fewer innovative offspring (Fig. S8), but leaves the patterns observed in Fig. 1 intact (Figs. S9, and S10).

In sum, each of the 50 carbon usage phenotypes we consider can give rise to metabolic innovations. Conversely, recombinants can acquire viability on each of 50 carbon sources. Thus, at least from this analysis, there is no evidence for absolute constraints on carbon usage phenotypes. However, different carbon usage phenotypes differ greatly in their propensity to arise as metabolic innovations, providing a first line of evidence for relative constraints on metabolic innovation by parental phenotypes.

**Novel metabolic phenotypes are relatively constrained by parental phenotypes**

Our next analysis goes to the heart of the question we pose. For each of the 50 focal carbon sources $C_i$, we examined all innovative offspring originating from parents viable on $C_i$ to find out whether gaining viability on each of the other 49 carbon sources ($C_j$, j $\neq i$) is possible. For 43 of the 50 carbon sources $C_i$, this is the case (Fig. 2a). That is, for such a parental carbon source $C_i$, at least one innovative offspring exists that gains viability on some new carbon source ($C_j$, j $\neq i$). Even for the remaining seven carbon sources $C_i$, this holds for the majority of the carbon sources $C_j$. That is, starting from viability on five of the seven carbon sources $C_i$, recombination can produce viability on more than 40 of the 49 carbon sources $C_j$. The remaining carbon sources $C_i$ are deoxyadenosine and adenosine, where recombination can produce metabolisms viable on 30 and 26 other carbon sources, respectively. Similar observations emerge when we repeat this analysis by increasing the number of reactions exchanged during recombination (Figs. S11a and S12a). In sum, for a majority (43 of 50) of parental phenotypes, there are no absolute constraints on metabolic innovation, i.e., all novel metabolic phenotypes considered here can arise through recombination.

Our next analysis (Fig. 2b) provides evidence for abundant relative constraints, that is, some carbon use phenotypes $C_j$ are more likely to emerge as metabolic innovations than others from parents viable on a given carbon source $C_i$. For example, 65.13% of the innovative offspring emerging from parents viable on glucose, gain viability on only 4 other carbon sources: 16.37% on D-fructose 6-phosphate, 17.72% on D-glucose 6-phosphate, 15.15% on D-fructose, and 15.89% on D-gluconate. The other 34.87% of metabolic innovations are distributed among 45 other carbon sources (on average each receiving 0.77% of the innovative offspring). As another example, for

parents viable on D-serine, 46% of the innovative offspring gain viability on glycine (9.71%), L-aspartate (11.4%), L-alanine (16.73%) or D-alanine (8.16%) and the rest of 54% innovations is distributed among the other 45 carbon sources (each on average 1.2%).

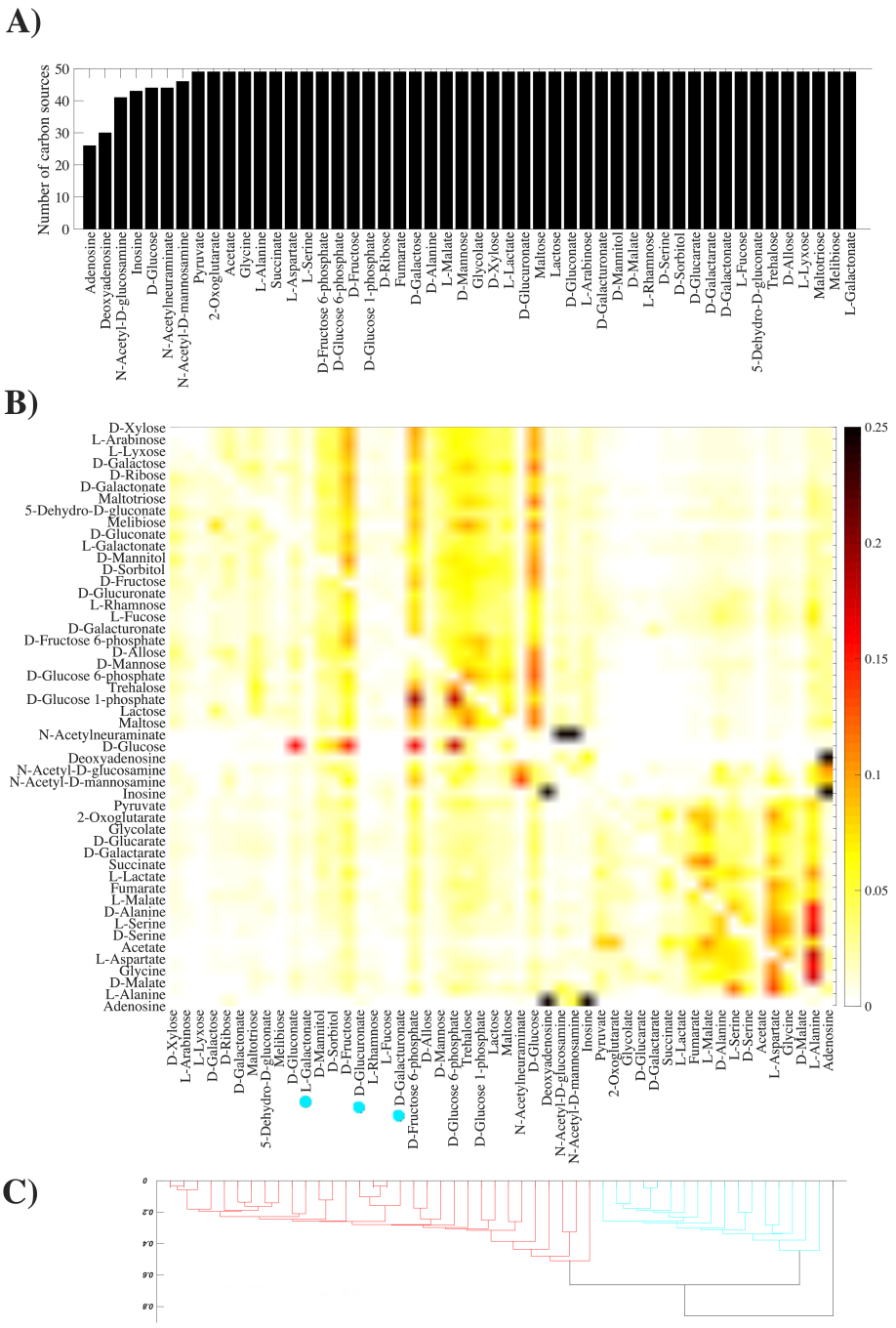



**Figure 2: Emergence of innovative offspring can be constrained by parental phenotypes. A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring recombination between parental metabolic networks is viable. **B)**

Fraction of innovative recombinants (coded according to the color legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, respectively (except D-galacturonate, L-galactonate, and D-glucoronate (cyan circles), which are the gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks during recombination.

We then clustered the 50 carbon sources based on their relative "innovation distance" in Fig. 2b, where two carbon sources $(C_i, C_j)$ are more distant if parents viable on $C_i$ give rise to fewer offspring viable on $C_j$. Fig. 2c shows that all glycolytic carbon sources (see S3 text) form one major branch of the resulting tree (colored red), and 17 of the 20 gluconeogenic carbon sources (except D-galacturonate, L-galactonate, D-glucoronate) form another major branch (colored cyan). Hence, the propensity for innovation between carbon sources belonging to the same class is higher than those belonging to different classes. This observation hints at a cause of the relative constraints we observe, which we discuss in more detail in section 3.4.

We observe qualitatively identical patterns when we repeat this analysis with altered numbers of reactions exchanged during recombination (Figs. S11 and S12), with altered genotypic distances among metabolic networks (Fig. S13), with smaller metabolic networks (Figs. S14 and S15) and with heterogeneous parental phenotypes (Figs. S16 and S17). However, in smaller metabolic networks, perhaps due to a substantially lower incidence of phenotypic innovation (Fig. S4), emergence of novel phenotypes is more constrained by parental phenotypes (Figs. S14 and S15). Moreover, for heterogeneous parental phenotypes where all the recipients are viable only on glucose and donors are viable on other carbon sources, carbon sources do not cluster according to innovation distance. The likely reason is that the recipient parental phenotype is constant in this analysis (Fig. S17).

In sum, different novel phenotypes are constrained in their evolution, because they originate with different probabilities from a given parental carbon-usage phenotype.

**Emergence of innovative offspring is not absolutely but relatively contingent on parental phenotypes**

To complement our above analyses, we also studied whether some novel metabolic phenotypes are *absolutely* contingent on a specific parental phenotype. That is, can they only emerge from parents with this phenotype? To find out, we studied the parental phenotypes of all innovative offspring that have gained viability on a given carbon source $C_j$, and did so for all carbon sources $C_j$. Fig. S18a shows that for all novel carbon-usage phenotypes $C_j$, innovative offspring can emerge from parents with at least 40 different phenotypes. Similar observations emerge when recombination alters a different number of reactions (Figs. S19a and S20a).

While absolute contingency does therefore not exist in our study system, we observe *relative* contingency: Different parental phenotypes $C_i$ have a greater or lesser propensity to give rise to a given carbon-usage phenotype $C_j$ (Fig. S18b).

For example, 42.15% innovative offspring gaining viability on D-galactarate originate from parents viable only on 4 different carbon sources, namely D-malate (12.86%), D-galacturonate (11.99%), pyruvate (8.70%), and glycolate (8.61%). The other 57.85% originate from parents viable on the other 45 carbon sources (where each accounts for 1.28% of the innovative offspring on average). Another example regards viability on succinate, 20.8% of which originates from parents viable on acetate and the rest is distributed among other parental phenotypes (each contributing 1.65% on average).

And once again, classification of carbon sources based on their distance (Fig. S18b) results in separation of glycolytic and gluconeogenic carbon sources (Fig. S18c). We observe similar patterns when we repeat this analysis with a different number of reactions altered during recombination (Figs. S19 and S20), with higher genotypic distances among parental metabolisms (Fig. S21), with smaller metabolic networks (Figs. S22 and S23), and with heterogeneous parental phenotypes (Figs. S24 and S25). In smaller metabolic networks, perhaps due to the lower incidence of innovation (Fig. S8), relative contingency is most pronounced (Figs. S22 and S23).

In sum, while we do not observe absolute contingency, some parental phenotypes are much more likely than others to give rise to specific new metabolic phenotypes, which show relative contingency.

**On the underlying causes of constraints and contingencies**

As we observed in Figs. 2c and S18c, one specific measure of biochemical similarity among carbon sources can help explain the patterns of constraints and contingencies that we observed. That is, carbon sources can be broadly partitioned into glycolytic and gluconeogenic classes, where parents viable on a carbon source in one class are most likely to produce innovative offspring viable on a new carbon source in the same class. To provide complementary evidence that constraints increase with biochemical distance among carbon sources, we used two other biochemical similarity measures, and determined whether they are associated with the innovation distance between carbon sources.

The first defines the metabolic distance between a given pair of carbon sources $(C_i, C_j)$ as the average shortest path between $C_i$ and $C_j$ in the substrate graph of 1,000 metabolic networks viable on $C_i$ (see supplementary text S7). This network-based biochemical distance is significantly associated with the number of recombinants that are generated from parents viable on $C_i$, and that gain viability on carbon source $C_j$ (Fig. S26, Pearson $r$ = -0.2722, and $P < 10^{-41}$). A second quantifier of distance relies on the superessentiality index, the proportion of random viable networks in which a given reaction is essential for viability on a given carbon source (see supplementary texts S7 and S8). Here also, innovation declines with increasing biochemical distance among carbon sources (Pearson $r$ = -0.3935, and $P < 10^{-83}$, supplementary text S8, Fig. S27a).

Another complementary analysis involving biochemical distance focuses on the individual reactions that can be transferred from donor to recipient, and that can lead to metabolic innovation. For this analysis, it is relevant that the majority of metabolic innovations is caused by the transfer of a single key reaction (32). We analyzed transferable reactions in greater depth, focusing on all 1,000 parental donor metabolic networks viable on a given carbon source $C_i$, and on the $(D/2 = 50)$ reactions that are present in the donor metabolic network, but are absent in the recipient, and so can potentially be transferred from the donor to the recipient. Specifically, we quantified the fraction of the 1,000 parental donor metabolic networks viable on $C_i$ in which at least one reaction among the $(D/2 = 50)$ transferrable reactions can have $C_j$ as a product or substrate, reasoning that such reactions may be especially prevalent among

reactions causing viability on $C_j$. The number of innovative offspring that gain viability on $C_j$ by recombining parents viable on $C_i$, increases significantly with the fraction of transferable reactions that involves $C_j$ (Pearson $r = 0.163$, and $P < 10^{-15}$, Fig. S27b). It is not difficult to see that this association can also be a consequence of the relatedness of two carbon sources. The reason is that metabolic networks viable on a given carbon source $C_i$ are likely to already have some reactions involving metabolically related carbon sources $C_j$. In this case, it is more likely that addition of a single novel reaction leads to the completion of a pathway in the recipient that is needed to metabolize $C_j$. We note that these correlation coefficients, albeit statistically significant, are low in magnitude, implying that these properties cannot fully explain the mechanism underlying phenotypic constraint. A more detailed analysis of each pathway connecting different carbon sources may be required to fully understand the causes of constraints and contingencies. We leave such an analysis for future work.

**Emergence of innovative offspring is constrained by and contingent on parental genotypes of both donors and recipients**

Our analyses thus far were focused on parental metabolic networks with given phenotypes, which allowed us to analyze constraints and contingencies emerging from such phenotypes. However, the emergence of novel phenotypes may also depend on parental genotypes, and we next analyzed such constraints. Random viable metabolisms are less than ideal for such an analysis for two reasons. First, they do not derive from any one organism with its specific gene-reaction association, and they do therefore not allow us to define genotypes on the level of genes. Second, our simple model of recombination for such metabolisms neglects the linkage of metabolic genes on chromosomes.

To overcome these limitations, we focused our next analysis on curated metabolic networks of 55 distinct bacterial strains or species. Their metabolic genes, reactions, gene-reaction association rules, metabolic gene locations, and biomass reactions are well-studied and available from the BiGG database (54). We used 30 carbon sources on which none of the 55 metabolisms are viable to study the emergence of novel phenotypes (supplementary text S3). We examined all 2970 (=55×54) distinct pairs of donor-recipient species or strains, and subjected them to recombination events that take into account metabolic gene linkage (see methods section 2.4). From each donor-

recipient pair, we generated millions of recombinant offspring to identify innovative offspring, that is, offspring gaining viability on at least one of the 30 novel carbon sources.

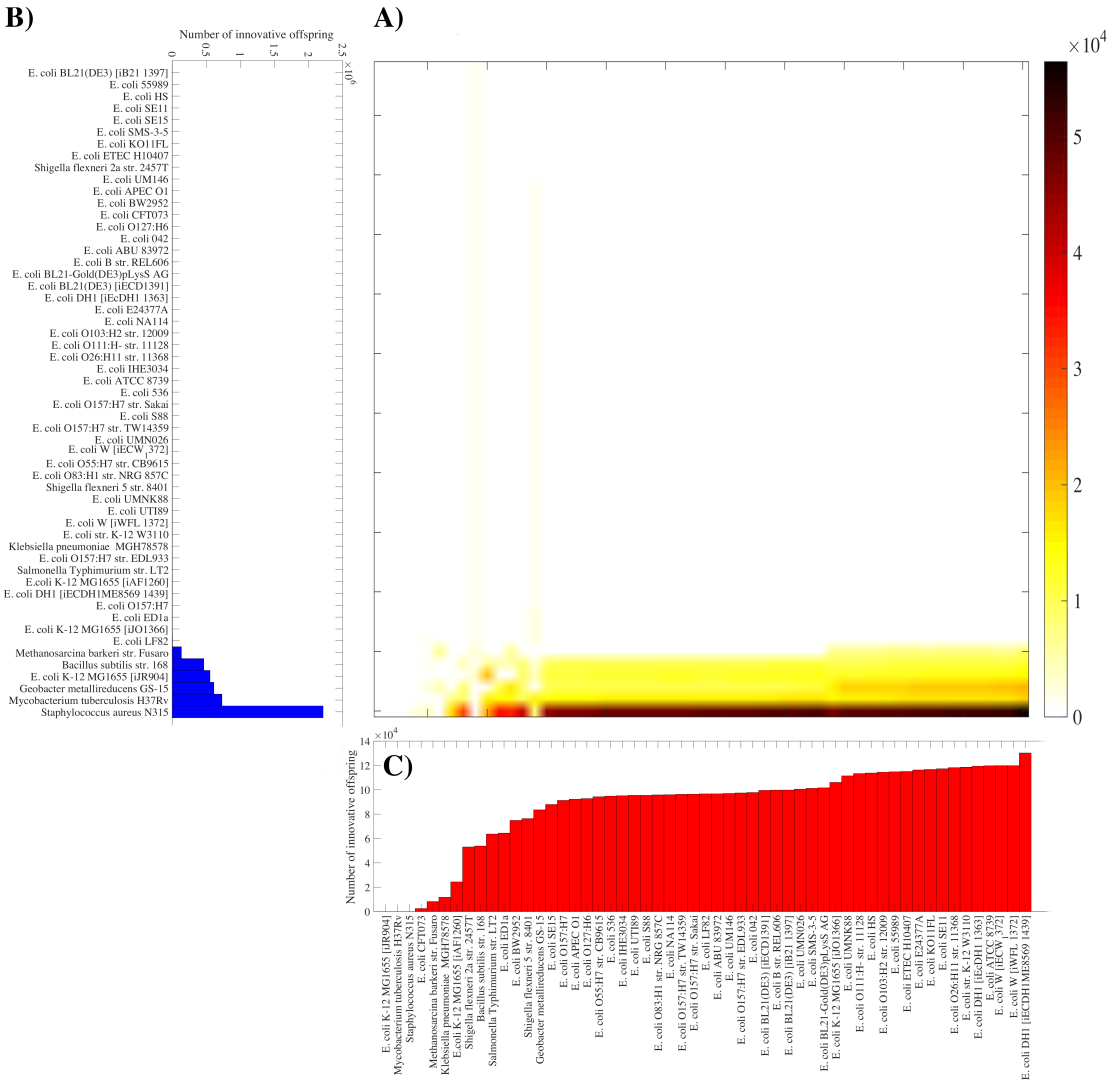



**Figure 3: Emergence of innovative offspring is contingent on and constrained by parental genotypes. A)** Number of innovative recombinant offspring resulting from linkage-based recombination between bacterial DNA donors specified on the vertical axis of panel B, and the corresponding recipients specified on the horizontal axis of panel C (number of recombinants encoded according to the color legend). **B)** Total number of innovative recombinant offspring involving the donor genotype specified on the vertical axis. **C)** Total number of innovative recombinant offspring involving the recipient genotype specified on the horizontal axis.

We observed that the emergence of novel phenotypes is strongly contingent on the recombining parental genotypes. Among the 2970 pairs of recombining parental genotypes, only 347 pairs (11.68%) brought forth at least one innovative offspring (Fig. 3a). In addition, these 347 pairs vary greatly in the number of innovative offspring that they can generate. The highest number of innovative offspring (56,461, or 1.17% of recombination events) emerges when the donor is *Staphylococcus aureus N315* and the recipient genotype is *E. coli DH1*, and the lowest number ((904), or 0.02% of recombination events) emerges when the donor is *E. coli BL21* and the recipient genotype is *Bacillus subtilis*.

The emergence of innovative offspring was also strongly constrained by the donor genotype (Fig. 3b). 97.84% of all innovative offspring identified in this analysis were generated from only six donor genotypes. The other 49 donors together were responsible only for 2.16% of all innovative offspring. Recombination involving *Staphylococcus aureus N315* donors caused an exceptionally large fraction of 45.97% of innovative offspring. Despite this strong relative constraint on donor genotypes, we did not observe any absolute constraints, because all 55 prokaryotic metabolisms generated at least one innovative offspring as donor genotypes, even though the contributions of 49 metabolisms were so small that they are not visible in Fig. 3b.

In contrast, the emergence of innovative offspring was not strongly constrained by the parental recipient genotype. That is, the majority of recipient metabolisms (48 out of 55) can generate approximately the same number of innovative offspring (Fig. 3c). Only four of them generated considerably fewer innovative offspring, and three of them did not generate any innovative offspring as recipients (Fig. 3c). Importantly, the potential of metabolic genotypes in generating innovative offspring when used as donors or recipients was highly asymmetric. For example, although *Staphylococcus aureus* and *Mycobacterium tuberclosis* accounted for most innovative offspring as donor genotypes, they did not generate any innovative offspring as recipient genotype. Similarly asymmetric biases emerged when we repeated the analysis with a recombination approach that does not take into account metabolic gene linkage, suggesting that such asymmetry is not caused by gene linkage but by the metabolic gene content of genomes (Fig. S28). In sum, the emergence of innovative offspring is strongly contingent on the genotypes of parental donor-recipient pairs, and especially on donor genotypes.

## 4.4. Discussion

In this study we systematically analyzed the prevalence of constraint and contingency for emerging novel phenotypes in complex metabolic systems. We did so by computationally emulating recombination among thousands of parental metabolic network pairs with specific phenotypes, and created millions of recombinant metabolic networks.

Overall, we observed little evidence for *absolute* constraints in the origin of novel phenotypes, i.e., metabolic networks with most carbon usage phenotypes can give rise to all 50 novel carbon usage phenotypes we consider here. However, there is ample evidence for *relative* constraints, that is, some carbon usage phenotypes are much more likely to arise relative to others from any one parental carbon usage phenotype.

Similarly, we observed no *absolute* contingency in the origin of novel phenotypes, i.e., recombinant metabolic networks with a given novel carbon usage phenotype can originate from all 50 parental phenotypes. In contrast, relative contingency is pervasive. That is, a given novel carbon usage phenotype is much more likely to originate from some parental phenotypes than others. Importantly, our observations remain qualitatively unchanged when we alter various properties of parental genotypes, such as their genotypic distance, which suggests that the different extents of constraints we observe may be an inherent property of metabolic systems.

We also analyzed the causes of constraints and contingencies, where several complementary analyses point to the importance of biochemical similarities among carbon source pairs $(C_i, C_j)$, where parents are viable on $C_i$, and recombinant offspring gain viability on $C_j$. First, if parents are viable on a carbon source that belongs to one of two major biochemical classes (glycolytic or gluconeogenic), then recombinant offspring tend to gain viability on a carbon source within the same class (Figs. 2c and S18c). Second, the smaller the number of reactions is that separate $C_i$ and $C_j$ in a metabolic network, the greater is the likelihood that offspring gain viability on $C_j$. Third, offspring gain viability on $C_j$ most often if a reaction transferred between donor and recipient involves $C_j$. This, in turn is most likely if the recipient already harbors some reactions necessary to metabolize $C_j$, and thus if catabolizing $C_i$ and $C_j$ involves similar reactions. Our analysis used carbon sources

that are not very heterogenous. Many of them, for example, are sugars that play important roles in central carbon metabolism. This biochemical similarity among carbon sources reduces constraints, and it may be responsible for the paucity of evidence for absolute constraints.

One strength of our approach is that it can address contingency and constraint in an entire class of system, and not just a single organism. However, the approach also has several limitations. First, any study relying on sampling is sensitive to sample size. For example, if we had analyzed only 100 parental metabolic networks and 100 recombinants per pair, we would not have observed any innovative offspring for most parental carbon usage phenotypes. Thus, we would have misleadingly concluded that absolute constraints are frequent in our study system. And even though we had generated a (computationally expensive) sample of one million offspring for each parental phenotype, we did see a small number of carbon sources showing evidence for absolute constraints. Such apparent absolute constraints may disappear at even higher sample sizes (Fig. S29a). In contrast, our assertion that relative constraints exist is less sensitive to sample sizes (Fig. S29b). Our current analysis generated fewer than 1,000 innovative metabolisms for most $(C_i, C_j)$ pairs, and larger sample sizes may help find out why some pairs $(C_i, C_j)$ are more or less involved in metabolic innovation.

Second, our work is based on flux balance analysis (15, 16), which neglects the influence of gene and enzyme regulation. However, because regulatory changes towards optimal expression of enzymes readily occur, even on the short time scales of laboratory evolution, this limitation may not affect our main observations (supplementary text S4).

Third, a recent study showed that the genome-scale metabolic networks are likely to include thermodynamically impossible energy-generating cycles (EGCs), which are capable of charging energy metabolites without nutrient consumption (56). These EGCs can artificially inflate biomass flux and so may mislead evolutionary simulations. Most of our randomly sampled viable metabolisms indeed harbor EGCs (97.3% and 97.8% of sampled metabolisms viable on glucose and acetate, respectively; Supplementary text S9). However, these EGCs do not strongly affect the emergence of novel phenotypes, nor do they substantially distort the patterns of relative constraint we observed (see supplementary text S9 and Figs. S30, S31, and

S32).

Finally, in our simulations using random metabolic networks, following common practice in the field (17–26), we define metabolic genotypes on the level of biochemical reactions rather than on that of genes or DNA. This representation neglects potentially important information, and especially the linkage of related metabolic genes on chromosomes, which may affect the outcome of recombination. To address this limitation, we also modeled recombination among metabolisms of 55 prokaryotic species or strains in a way that includes gene linkage information. This analysis also demonstrates strong constraints and contingencies in the emergence of novel metabolic phenotypes.

A previous experimental evolution study suggested a strong relative constraint in the emergence of a novel citrate utilization phenotype, which required thousands of generations of laboratory evolution subject to mutation and selection to emerge (9). Although our simulations are not strictly commensurate with any experimental study, for example because we do not consider DNA changes explicitly, we speculate that such relative constraints would be less pronounced in any system where recombination is abundant, because recombination can cause larger scale changes than mere point mutations that would alter individual reactions or transport processes (9).

This was one motivation to choose recombination as an agent of genetic change in the first place, reasoning that any constraints visible in the presence of recombination might be even stronger in the presence of less dramatic genetic changes.

Metabolic systems are one of the three classes of biological systems in which phenotypic variation is crucial for evolutionary adaptation and innovation (57). The other two are macromolecules (protein and RNA) and regulatory systems. Predicting phenotypes in these systems is less straightforward than for metabolic systems (58–60). In proteins, for example, phenotypes form through a complex and incompletely understood three-dimensional folding process (58), and in regulatory systems, gene expression phenotypes emerge from complex interactions among regulatory molecules (59, 60). Our understanding of inherent biases in phenotypic variability will not be complete until we understand contingencies and constraints in these classes of systems as well, which remains an important task for future work.

## 4.5. Methods

### Genotype-phenotype representation in metabolic networks

The set of enzyme-catalyzed biochemical reactions that take place in an organism constitutes the organism's metabolic reaction network, i.e., its metabolism. Each such metabolism contains a subset of the "reaction universe" of all biochemical reactions that are known to occur in some organism within the biosphere. We have manually curated a representation of the prokaryotic reaction universe, which comprises 5,906 reactions known to occur in prokaryotes (see supplementary texts S1 and S2 for details). In this framework, we represent an organism's metabolic genotype as a binary vector of length 5,906, each entry of which corresponds to a given reaction in the universe, and is equal to one if the corresponding reaction is present in the network, and zero otherwise. Hence, each genotype can be envisioned as a single member of a vast space of all possible metabolic networks, which contains $2^{5906}$ distinct genotypes. We determine the phenotype of a given metabolic genotype based on its ability to sustain life in one or more of 50 distinct minimal environments that differ only in the sole carbon source they contain (supplementary text S3). We consider a genotype *viable* on a given carbon source, if Flux Balance Analysis (FBA, See supplementary text S4) predicts that it can produce all essential biomass precursors using this carbon source as its only carbon source (15). We used the biomass composition of the *E. coli* metabolic model iAF1260, because the sampling approach described in the next section starts from the *E. coli* metabolism (supplementary text S5). Our C++ implementation of FBA and the code necessary for the analyses in this paper are available through this public github repository: https://github.com/rzgar/EMETNET.

### Random sampling of parental metabolic network pairs from metabolic genotype space

We here employ a previously described *in silico* process that relies on Markov Chain Monte Carlo (MCMC) random walks to generate randomly sampled viable metabolic networks, i.e., networks that are viable on a given carbon source, but that otherwise contain a random subset of reactions in the reaction universe (supplementary text S5) (21, 23). This procedure ensures uniform sampling from the set of all metabolic

networks viable on a given carbon source. Our analyses required us to recombine pairs of "parental" metabolic networks (i.e., donor-recipient pairs) with particular features, such as a given genotypic distance (*D*), defined as the number of reactions differing between the parents. We used simultaneous genotype-converging MCMC random walks to generate pairs of metabolic networks with a given *D* (See supplementary text S6). We required parental metabolisms to be exclusively viable on a particular carbon source, i.e., to be inviable on all 49 other carbon sources we considered. In most of our analyses, we kept the number of reactions present in the metabolic networks constant and equal to that of *E. coli* with 2,079 reactions.

**Modeling a recombination-like process in metabolic networks**

As in a previous contribution (32), we use a coarse-grained model of prokaryotic recombination that mimics the effects of horizontal gene transfer events between bacteria on metabolism (33–36). This model is motivated by the importance of horizontal gene transfer as a means of genetic change. Through its high incidence, horizontal gene transfer can change the gene content of genomes on short evolutionary time scales (33, 37, 38). It can also occur between very distantly related organisms (39, 40). For several reasons, our recombination model also takes DNA deletions into consideration. The first is that during horizontal gene transfer, incorporating genes from a donor into a recipient genome relies on DNA rearrangements that can also delete resident genes (41). Second, the majority of newly acquired genes obtained via horizontal gene transfer reside in the genome only for short amounts of time (42). Third, the evolution of prokaryotic genomes is biased towards DNA deletions (43). Motivated by these observations, we here model prokaryotic recombination as a process where the transfer of biochemical reactions from a donor to a recipient is accompanied by concurrent deletion of reactions from the recipient metabolic network.

Specifically, to model recombination for each parental metabolic network pair, we generated 1,000 recombinant offspring by (i) adding to the recipient metabolic network a given number *n/2* of randomly chosen reactions that were present in the donor and absent from the recipient, followed by (ii) deleting *n/2* reactions randomly chosen from the recipient. Thus, the total number of reactions changed by a recombination event in the recipient is equal to *n*. In this present contribution, we repeated most of our analyses by using three different values of *n*; namely $n = 10, 20,$

and 30. Empirical observations also suggest that altering up to $n = 60$ reactions in a recombination event is biologically realistic, because horizontal gene transfer can affect long DNA regions (44). Importantly, the transferred material that is integrated into the host genome by recombination can constitute stretches of non-coding DNA, fragments of genes (45, 46), entire genes (47), multiple adjacent genes (48, 49), operons, transposable chromosomal elements, plasmids, as well as other naturally occurring extrachromosomal elements (50). The length of contiguously transferred stretches may range from a few nucleotides (51) to more than 3 Mbp (44), i.e., some two thirds of the length of the *E. coli* genome, which encodes more than 1300 reactions. In addition, some megabase-scale horizontally transferred DNA segments can become incorporated into a chromosome in the form of hundreds of smaller fragments (52). As we have discussed in a previous contribution ((32), electronic supplementary text S3), the probability that a recombination event preserves viability exceeds $10^{-3}$ for values up to $n = 60$.

**Modeling recombination in curated bacterial metabolic networks from the BiGG database**

We used the R-package Sybil (53) to collect 55 well-annotated bacterial genome-scale metabolic networks available in the BiGG database (54). Each of these species or strains has its own biomass growth function, its own complement of reactions, and well defined gene-reaction association rules that allowed us to model recombination on the level of genes instead of reactions. We used the genomic location of metabolic genes in these bacterial species or strains (55) to take gene linkage into account when modeling recombination.

To generate a recombinant metabolic network from a donor and a recipient organism, first a given stretch of DNA from the donor genome that contains a given number of metabolic genes is translated into reactions based on the gene-reaction association rules of the donor organism, and then the resulting reactions are added to the recipient metabolic network. Second, a given stretch of DNA from the recipient genome that contains a given number of metabolic genes is translated into reactions based on the gene-reaction association rules of the recipient organism, and then the resulting reactions are deleted from the recipient metabolic network.

In a recombination event between a pair of organisms, we set the number of genes in a given donor DNA stretch such that on average a given number of *n=5* reactions are added to the recipient metabolic network, and on average an equal number *n=5* of reactions are deleted from it. Because gene-reaction associations are not generally one-to-one and can be very complicated, and because most of the reactions that are encoded in a given stretch of DNA may already be present in the recipient metabolic network, the number of metabolic genes that needs to be added from donor to recipient genome, such that exactly *n* reactions are added to the recipient, will often be higher than *n*. In contrast, we found that the number of metabolic genes in a DNA stretch to be deleted from the recipient genome in order to eliminate *n* reactions from the recipient metabolic network is lower than *n*, because deletion of a single metabolic gene often causes elimination of multiple reactions.

More specifically, we modeled recombination among all distinct pairs of donor-recipient bacterial species or strains in our analysis ($(55\times54)$ pairs). From each given pair we generated a recombinant offspring by adding a given ($p$) number of consecutive metabolic genes from the donor genome, followed by deleting a given ($q$) number of consecutive metabolic genes from the recipient genome. Importantly, we examined all possible combinations of ($p$) consecutive genes from the donor and ($q$) consecutive genes from the recipient. Thus, for a donor genome with $n$ metabolic genes, and a recipient genome with $m$ metabolic genes, we generated all $(n-p+1)\times(m-q+1)$ recombinant offspring, a number that exceeded one million offspring for most pairs. Note that ($p$) and ($q$) are selected based on the gene-reaction association rules of the donor and recipient species or strains to ensure that any one recombination event adds on average 5 new reactions and deletes 5 reactions from the recipient metabolic network.

To study the effect of linkage on the emergence of novel phenotypes, we followed a second recombination procedure that neglects linkage between metabolic genes. That is, we added or deleted reactions randomly, just as we had done for randomly sampled metabolic networks, irrespective of the genomic position of the metabolic genes encoding these reactions. To do so, we examined all distinct donor-recipient pairs, and from each pair we generated the same number ($(n-p+1)\times(m-q+1)$) of recombinant offspring as in the linkage-based approach, ensuring that on average 5

randomly chosen reactions are added from the donor and deleted from the recipient metabolic network.

To identify innovative offspring among all the generated recombinants, we used 30 carbon-containing metabolites on which none of the 55 bacterial species or strains are predicted to be viable (listed in supplementary text S3). To predict viability of a recombinant metabolic network using FBA, we used the objective function of the recipient, because recombinants are much more similar to the recipient than to the donor.

## 4.6. References


1. Maynard-Smith, J., R. Burian, S. A. Kauffman, P. Alberch, J. Campbell, B. Goodwin, R. Lande, D.R. and L.W. 1985. Developmental Constraints and Evolution. Q. Rev. Biol. 60: 265–287.
2. Wagner, A. 2011. Genotype networks shed light on evolutionary constraints. Trends Ecol. Evol. 26: 577–584.
3. Nüsslein-Volhard, C., and E. Wieschaus. 1980. Mutations affecting segment number and polarity in Drosophila. Nature. 287: 795–801.
4. West, G.B., J.H. Brown, and B.J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. Science. 276: 122–6.
5. Nelson, D.L., and M.M. Cox. 2004. Lehninger Principles of Biochemistry. 3rd ed. New York: W. H. Freeman.
6. Levitt, M. 2009. Nature of the protein universe. Proc. Natl. Acad. Sci. U. S. A. 106: 11079–84.
7. Stephen Jay Gould. 1990. Wonderful Life: The Burgess Shale and the Nature of History. New York: W. W. Norton & Company .
8. Lobkovsky, A.E., and E. V Koonin. 2012. Replaying the tape of life: quantification of the predictability of evolution. Front. Genet. 3: 246.
9. Blount, Z.D., C.Z. Borland, and R.E. Lenski. 2008. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. 105: 7899–906.
10. Copley, S.D. 2000. Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. Trends Biochem. Sci. 25: 261–5.
11. Rehmann, L., and A.J. Daugulis. 2008. Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. Biotechnol. Bioeng. 99: 521–8.
12. van der Meer JR, C. Werlen, S. Nishino, and J. Spain. 1998. Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. Appl. Environ. Microbiol. 64: 4185–93.

13. Cline, R.E., R.H. Hill, D.L. Phillips, and L.L. Needham. Pentachlorophenol measurements in body fluids of people in log homes and workplaces. Arch. Environ. Contam. Toxicol. 18: 475–81.

14. Dantas, G., M.O.A. Sommer, R.D. Oluwasegun, and G.M. Church. 2008. Bacteria subsisting on antibiotics. Science. 320: 100–3.

15. Orth, J.D., I. Thiele, and B.Ø. Palsson. 2010. What is flux balance analysis? Nat. Biotechnol. 28: 245–8.

16. Edwards, J.S., R.U. Ibarra, and B.O. Palsson. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19: 125–30.

17. Edwards, J.S., and B.O. Palsson. 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc. Natl. Acad. Sci. U. S. A. 97: 5528–33.

18. Feist, A.M., and B.Ø. Palsson. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat. Biotechnol. 26: 659–67.

19. McCloskey, D., B.Ø. Palsson, and A.M. Feist. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol. 9: 661.

20. Lewis, N.E., H. Nagarajan, and B.O. Palsson. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. 10: 291–305.

21. Matias Rodrigues, J.F., and A. Wagner. 2009. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput. Biol. 5: e1000613.

22. Edwards, J.S., and B.O. Palsson. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274: 17410–6.

23. Samal, A., J.F. Matias Rodrigues, J. Jost, O.C. Martin, and A. Wagner. 2010. Genotype networks in metabolic reaction spaces. BMC Syst. Biol. 4: 30.

24. Barve, A., S.-R. Hosseini, O.C. Martin, and A. Wagner. 2014. Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. BMC Syst. Biol. 8: 48.

25. Hosseini, S.-R., A. Barve, and A. Wagner. 2015. Exhaustive Analysis of a Genotype Space Comprising 1015 Central Carbon Metabolisms Reveals an Organization Conducive to Metabolic Innovation. PLoS Comput. Biol. 11: e1004329.

26. Hosseini, S.-R., and A. Wagner. 2016. The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. BMC Syst. Biol. 10: 97.

27. Stemmer, W.P. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. Proc. Natl. Acad. Sci. 91: 10747–10751.

28. Zhang, Y.-X., K. Perry, V.A. Vinci, K. Powell, W.P.C. Stemmer, and S.B. del Cardayré. 2002. Genome shuffling leads to rapid phenotypic improvement in

bacteria. Nature. 415: 644–6.

29. Crameri, A., G. Dawes, E. Rodriguez, S. Silver, and W.P. Stemmer. 1997. Molecular evolution of an arsenate detoxification pathway by DNA shuffling. Nat. Biotechnol. 15: 436–8.

30. Chang, C.C., T.T. Chen, B.W. Cox, G.N. Dawes, W.P. Stemmer, J. Punnonen, and P.A. Patten. 1999. Evolution of a cytokine using DNA family shuffling. Nat. Biotechnol. 17: 793–7.

31. Ness, J.E., M. Welch, L. Giver, M. Bueno, J.R. Cherry, T. V Borchert, W.P. Stemmer, and J. Minshull. 1999. DNA shuffling of subgenomic sequences of subtilisin. Nat. Biotechnol. 17: 893–6.

32. Hosseini, S.-R., O. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc. R. Soc. B. .

33. Thomas, C.M., and K.M. Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3: 711–21.

34. Guttman, D.S., and D.E. Dykhuizen. 1994. Clonal divergence in Escherichia coli as a result of recombination, not mutation. Science. 266: 1380–3.

35. Feil, E.J., E.C. Holmes, D.E. Bessen, M.S. Chan, N.P. Day, M.C. Enright, R. Goldstein, D.W. Hood, A. Kalia, C.E. Moore, J. Zhou, and B.G. Spratt. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. U. S. A. 98: 182–7.

36. Whitaker, R.J., D.W. Grogan, and J.W. Taylor. 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon Sulfolobus islandicus. Mol. Biol. Evol. 22: 2354–61.

37. Ochman, H., J.G. Lawrence, and E.A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405: 299–304.

38. Pál, C., B. Papp, and M.J. Lercher. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat. Genet. 37: 1372–5.

39. Fraser, C., W.P. Hanage, and B.G. Spratt. 2007. Recombination and the nature of bacterial speciation. Science. 315: 476–80.

40. Majewski, J., P. Zawadzki, P. Pickerill, F.M. Cohan, and C.G. Dowson. 2000. Barriers to genetic exchange between bacterial species: Streptococcus pneumoniae transformation. J. Bacteriol. 182: 1016–23.

41. Kowalczykowski, S.C., D.A. Dixon, A.K. Eggleston, S.D. Lauder, and W.M. Rehrauer. 1994. Biochemistry of homologous recombination in Escherichia coli. Microbiol. Rev. 58: 401–65.

42. Kuo, C.-H., and H. Ochman. 2009. The fate of new bacterial genes. FEMS Microbiol. Rev. 33: 38–43.

43. Mira, A., H. Ochman, and N.A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17: 589–96.

44. Lin, C.H., G. Bourque, and P. Tan. 2008. A comparative synteny map of Burkholderia species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. Mol. Biol. Evol. 25: 549–58.

45. Bork, P., and R.F. Doolittle. 1992. Proposed acquisition of an animal protein domain by bacteria. Proc. Natl. Acad. Sci. 89: 8990–8994.

46. Inagaki, Y., E. Susko, and A.J. Roger. 2006. Recombination between elongation factor 1  genes from distantly related archaeal lineages. Proc. Natl. Acad. Sci. 103: 4528–4533.

47. Hartl, D.L., E.R. Lozovskaya, and J.G. Lawrence. 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica. 86: 47–53.

48. Igarashi, N., J. Harada, S. Nagashima, K. Matsuura, K. Shimada, and K. V Nagashima. 2001. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. J. Mol. Evol. 52: 333–41.

49. Omelchenko, M. V, K.S. Makarova, Y.I. Wolf, I.B. Rogozin, and E. V Koonin. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol. 4: R55.

50. Chan, C.X., R.G. Beiko, A.E. Darling, and M.A. Ragan. 2010. Lateral Transfer of Genes and Gene Fragments in Prokaryotes. Genome Biol. Evol. 1: 429–438.

51. Denamur, E., G. Lecointre, P. Darlu, O. Tenaillon, C. Acquaviva, C. Sayada, I. Sunjevaric, R. Rothstein, J. Elion, F. Taddei, M. Radman, and I. Matic. 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 103: 711–21.

52. Didelot, X., M. Achtman, J. Parkhill, N.R. Thomson, and D. Falush. 2007. A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res. 17: 61–8.

53. Gelius-Dietrich, G., A.A. Desouki, C.J. Fritzemeier, and M.J. Lercher. 2013. Sybil--efficient constraint-based modelling in R. BMC Syst. Biol. 7: 125.

54. King, Z.A., J. Lu, A. Dräger, P. Miller, S. Federowicz, J.A. Lerman, A. Ebrahim, B.O. Palsson, and N.E. Lewis. 2015. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. : gkv1049-.

55. Tatusova, T., S. Ciufo, B. Fedorov, K. O'Neill, and I. Tolstoy. 2014. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 42: D553-9.

56. Fritzemeier, C.J., D. Hartleb, B. Szappanos, B. Papp, M.J. Lercher, and G. Fekete. 2017. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Comput. Biol. 13: e1005494.

57. Wagner, A. 2011. The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems. Oxford: Oxford University Press.

58. Dill, K.A., S.B. Ozkan, M.S. Shell, and T.R. Weikl. 2008. The protein folding problem. Annu. Rev. Biophys. 37: 289–316.

59. Karlebach, G., and R. Shamir. 2008. Modelling and analysis of gene regulatory networks. Nat. Rev. Mol. Cell Biol. 9: 770–780.

60. De Smet, R., and K. Marchal. 2010. Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. 8: 717.

61. Goto, S., T. Nishioka, and M. Kanehisa. 2000. LIGAND: chemical database of enzyme reactions. Nucleic Acids Res. 28: 380–2.

62. Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 30: 402–4.

63. Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38: D355-60.

64. Feist, A.M., C.S. Henry, J.L. Reed, M. Krummenacker, A.R. Joyce, P.D. Karp, L.J. Broadbelt, V. Hatzimanikatis, and B.Ø. Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3: 121.

65. Lercher, M.J., and C. Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25: 559–67.

66. Ibarra, R.U., J.S. Edwards, and B.O. Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 420: 186–9.

67. Vieira-Silva, S., and E.P.C. Rocha. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 6: e1000808.

68. Kirschner, D., and S. Marino. 2005. Mycobacterium tuberculosis as viewed through a computer. Trends Microbiol. 13: 206–11.

69. Fong, S.S., and B.Ø. Palsson. 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat. Genet. 36: 1056–8.

70. Fong, S.S., J.Y. Marciniak, and B.O. Palsson. 2003. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J. Bacteriol. 185: 6400–6408.

71. Wagner, A., and D.A. Fell. 2001. The small world inside large metabolic networks. Proc. R. Soc. B Biol. Sci. 268: 1803–1810.

72. Barve, A., J.F.M. Rodrigues, and A. Wagner. 2012. Superessential reactions in metabolic networks. Proc. Natl. Acad. Sci. U. S. A. 109: E1121-30.

73. Ma, H.-W., and A.-P. Zeng. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19: 1423–30.

74. Hopcroft, J., and R. Tarjan. 1973. Algorithm 447: efficient algorithms for graph manipulation. Commun. ACM. 16: 372–378.

## 4.7. Supplementary Information

### S1: Genome-scale metabolic networks and their phenotypic representations

Similar to our previous work describing the procedures used here (1), and following common practice in metabolic systems biology (2–4), we represent an organism's metabolic genotype as the set of genomically encoded (enzyme-catalyzed) biochemical reactions proceeding inside the organism. This metabolic genotype specifies a metabolism or metabolic network, a network of chemical reactions encoded by the genotype. A metabolic reaction network enables an organism to extract energy and produce small biomass building blocks, such as amino acids, from extracellular nutrients. Inference of this genotype from genomic and biochemical information has been successful for multiple organisms (5, 6).

Any one metabolic reaction network contains a subset of the "reaction universe" of all biochemical reactions that take place in prokaryotes (See text S2). We have curated a representation of this universe, which comprises 5,906 reactions and is based on current metabolic knowledge (7–10). We represent an organism's metabolic genotype as a binary vector of length 5,906. Each entry of this vector corresponds to a given reaction in the reaction universe, and is equal to one if the corresponding reaction is present in the metabolic network, and zero otherwise. Thus, each genotype can be thought of as a single member of a vast space of all possible metabolic networks, which contains $2^{5906}$ distinct genotypes.

We define the phenotype of a given metabolic genotype based on its viability in 50 distinct minimal environments that differ only in the carbon source they harbor (See Text S3). We consider that a genotype is *viable* on a given carbon source, if it can produce all essential biomass precursor molecules from the given carbon source, and we use Flux Balance Analysis (FBA, See text S4) to determine viability (11). We represent the phenotype of a given metabolic genotype as a binary vector of length 50. Each entry of this vector corresponds to a given carbon source, and it is equal to one if the genotype is viable on this carbon source, and zero otherwise.

### S2: Reaction universe

The reaction universe we curated is a set of metabolic reactions in which each reaction is known to occur in some prokaryotic organisms. For the curation of this universe,

we used data from the LIGAND database (7, 8) of the Kyoto Encyclopedia of Genes and Genomes (9). Briefly, the LIGAND database, which is comprised of the REACTION and the COMPOUND databases, provides information on reactions, associated stoichiometric information, chemical compounds involved in a reaction, and the Enzyme Classification (E.C.) identifier of each reaction. From the REACTION and the COMPOUND databases we excluded (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and, (iv) reactions involving complex metabolites without chemical information about their structure (10). Moreover, we do not consider unknown reactions, and we also do not take into account spontaneous reactions, or reactions that depend on external stimuli. The published *E. coli* metabolic model (iAF1260) consists of 1397 non-transport reactions (12). We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the unique (non-duplicate) reactions. This resulted in a universe of reactions consisting of 682 transport, 5,906 non-transport reactions and 5030 metabolites. The reaction universe is available online (https://github.com/rzgar/EMETNET/tree/master/UNIVERSE).

**S3: Chemical environments**

We consider 50 minimal growth environments, each of which includes oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron ($Fe^{2+}$ and $Fe^{3+}$), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese, zinc, and a specific carbon source. Importantly, to represent different chemical environments, we vary the carbon source while keeping all other nutrients constant. We consider a metabolic network viable on a given carbon source, if it can synthesize all essential biochemical precursors when this carbon source is provided as the sole carbon source in the minimal medium just described.

We used 50 carbon sources for our analysis of randomly sampled metabolic networks, including the following 27 glycolytic carbon sources: D-glucose, D-glucose 6-phosphate, trehalose, maltose, lactose, D-fructose 6-phosphate, D-fructose, D-mannose, D-mannitol, D-glucose 1-phosphate, D-sorbitol, maltotriose, D-allose, D-ribose, D-xylose, D-gluconate, 5-dehydro-D-gluconate, L-rhamnose, L-fucose, L-

arabinose, L-lyxose, D-galactose, melibiose, D-galactonate, N-acetyl-D-glucosamine, N-acetyl-D-mannosamine, N-acetylneuraminate.

In addition, we used the following 20 gluconeogenic carbon sources: pyruvate, L-alanine, L-lactate, D-alanine, D-malate, acetate, L-serine, L-malate, D-serine, glycine, glycolate, L-aspartate, succinate, fumarate, 2-oxoglutarate, D-galacturonate, D-galactarate, D-glucarate, L-galactonate, D-glucoronate. And we used the following three nucleosides as carbon sources: adenosine, deoxyadenosine, inosine.

To study the emergence of novel phenotypes in 55 prokaryotic metabolic networks from the BiGG database (13) (see methods section 2.4 in the main text), we used the following 30 carbon sources on which none of the 55 metabolic networks are predicted to be viable: Biotin, riboflavin, folate, pimelate, urea, carbonic acid, bicarbonate, methanol, trimethylamine, D-methionine, glycine betaine, gamma-butyrobetaine, choline, L-phenylalanine, L-leucine, L-tyrosine, L-methionine, thiamin, 6-diaminoheptanedioate, (R)-pantothenate, spermidine, taurine, isocytosine, protoheme, nicotinamide adenine dinucleotide, L-fucose 1-phosphate, dimethyl-sulfide, L-carnitine, dimethyl sulfoxide, and 1,5-diaminopentane.

**S4: Flux balance analysis**

Flux balance analysis (FBA) is a computational method that is widely used for the quantitative analysis and modeling of metabolic networks (11). Based on the stoichiometric coefficients of the metabolites participating in the reactions of a given metabolic network, FBA predicts the metabolic flux through each reaction. Stoichiometric coefficients are stored in a stoichiometric matrix *S*, which is of dimension $m \times n$, where $m$ and $n$, denote the number of metabolites and the number of reactions in a metabolic network. FBA constrains the flux through each reaction based on the assumption that a metabolic network is in a steady state where metabolite concentrations do not change, i.e., $Sv = 0$, where $v$ is the vector of metabolic fluxes $v_i$ through reaction $i$. The solutions of the equation $Sv = 0$, that is, the null space of matrix S, comprises all flux vectors that are allowable in steady state. The null space is further constrained by physicochemical information regarding the maximum and minimum possible fluxes through each reaction. FBA relies on an optimization procedure called linear programming to identify those among the allowable flux

vector(s) that maximize an objective function *Z*. This task can be formulated as finding a flux vector $v^*$ with the property

$$v^* = \max_v Z(v) = \max_v \{ c^T v \mid Sv = 0,\ a \leq v \leq b \},$$

where the vector *c* contains a set of scalar coefficients representing the maximization criterion, and each entry $a_i$ and $b_i$ of vectors $a$ and $b$, indicates the minimally and maximally possible flux through reaction *i*. The vector *c* represents the proportions of each small biomass molecule in a cell's biomass. Therefore $v^*$ maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass (11). Here we use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and we consider a metabolic network viable if it can produce all essential biomass precursors. More precisely, FBA predicts a metabolic network as viable on a given environment, if its biomass flux rate exceeds 0.001 $1/h$. In a free-living bacterium like *E. coli,* there are approximately 60 such molecules including 20 amino acids, DNA, and RNA precursors, lipids and cofactors. We used the biomass composition of the *E. coli* metabolic model iAF1260 to define the vector *c* (12). Moreover, we used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problem of FBA.

The major limitation of FBA is that it neglects regulatory constraints that can arise through suboptimal expression or regulation of enzymes. Newly horizontally transferred genes cannot easily establish regulatory interactions with their host genes, and it may thus take considerable adaptive evolution until they become expressed at a maximal or optimal level (14). Such regulatory constraints would be especially important if we focused on quantitative predictions of biomass growth (15). However, we use FBA solely for qualitative prediction of viability. This focus on qualitative phenotypes is biologically sensible. The reason is that many organisms grow slowly in their native environment (16, 17), implying that regulation for maximal biomass production is far from universal. Moreover, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments (15, 18, 19).

**S5: Generation of random metabolic networks**

We here employ a previously described *in silico* process which relies on Markov Chain Monte Carlo (MCMC) random walks to generate metabolic networks that comprise random sets of metabolic reactions that are viable on a given carbon source (10, 20). This procedure can produce metabolic networks that are sampled uniformly from the set of all metabolic networks viable on a given carbon source (10, 20). Briefly, in each step of such a random walk we perform a reaction swap, defined as altering a metabolic network by adding a randomly chosen reaction from the reaction universe, and then deleting a reaction randomly chosen from the set of reactions present in the metabolic network. If the reaction swap disrupts the metabolic network's viability on the given carbon source (as determined by FBA) we reject it, and perform another reaction swap until we find a swap that does not disrupt viability. This procedure also ensures that the total number of reactions remains constant. For the MCMC method to produce random samples of metabolic networks, it is essential to carry out enough reaction swaps to "erase" the random walker's similarity to the initial metabolic network. Previously, it has been shown that $3 \times 10^3$ reaction swaps are sufficient for this purpose (10, 20). Each of our random walks starts from *E. coli*'s metabolic network and performs $10^4$ reaction swaps before storing the final metabolic network for further analysis. We used $10^4$ independent random walks conducted in this way to create $10^4$ random metabolic networks viable on each of the 50 carbon sources.

**S6: Generation of parental metabolic network pairs**

Some of our analyses required us to recombine pairs of "parental" metabolic networks with particular features, such as being viable on a specific carbon source (and only on that carbon source), or having a given genotypic distance ($D$), defined as the number of reactions differing between the parents. Generating parents with a given genotypic distance ($D$) is not straightforward, because the random metabolic networks generated by MCMC sampling generally have genotypic distances sufficiently large ($D \approx 2{,}000$) to be biologically unrealistic for modeling frequently recombining prokaryotic genomes. To create less distant metabolic network pairs, we took an MCMC random walk approach. It revolves around a reaction-swapping random walk starting with a pair of randomly chosen metabolic networks from our sample of $10^4$ sampled

metabolic networks that are exclusively viable on a given carbon source. In each step of this random walk, we subjected each parental metabolic network to a reaction swap, and we accepted each reaction swap if it (*i*) preserved the original phenotype, and (*ii*) did not increase the genotypic distance of the two metabolic networks after the swap, otherwise we rejected the reaction swap. We continued this procedure until the genotypic distance between the metabolic networks became equal to a desired distance *D*. We note that this procedure is very time-consuming when applied to the thousands of parents we study here.

Finally, to generate parental metabolic networks with a given number of reactions, we started from a random viable metabolic network generated by MCMC sampling, as described in the text S5. All such metabolic networks have the same number of reactions as *E. coli* (2,079). We then applied a sequence of individual and random reaction deletions, where we required that each deletion preserve viability, until the network had reached the desired size.

**S7. Estimation of the metabolic distance between carbon sources**

For each pair of carbon sources $(C_i, C_j)$, we calculated metabolic distance with two different approaches, a direct approach that is based on the shortest path between carbon sources in substrate graph (21), and an indirect approach that is based on carbon source-dependent superessentiality of metabolic reactions in metabolic networks (22).

The first approach relies on the substrate graph of a metabolic network, in which vertices correspond to metabolites. Two metabolites are linked via an edge, if the metabolites participate in the same metabolic reactions as either a substrate or a product. From this substrate graph we excluded currency metabolites, which are metabolites that transfer small chemical groups, and are involved in many reactions (23). Specifically, we excluded protons, $H_2O$, ATP (adenosine triphosphate), ADP (adenosine diphosphate), AMP (adenosine monophosphate), NADP(H) (nicotinamide adenosine dinucleotide diphosphate), NAD(H) (nicotinamide adenosine dinucleotide), and $P_i$ (inorganic phosphate), CoA (coenzyme A), hydrogen peroxide, ammonia, ammonium, bicarbonate, GTP (guanosine triphosphate), GDP (guanosine diphosphate), and $PP_i$ (inorganic diphosphate) that occurred in both the cytoplasmic and periplasmic compartments. In addition, we excluded oxidized and reduced forms of cofactors such as quinone, ubiquinone, glutathione, thioredoxin, flavodoxin and

flavin mononucleotide. For all metabolic networks viable on $C_i$, we measured the shortest path in the substrate graph between $C_i$ and any other $C_j, j \neq i$ using Dijkstra's algorithm (24). Then, we considered the average shortest path between $C_i$ and $C_j$ among metabolic networks viable on $C_i$ as the metabolic distance between $C_i$ and $C_j$.

In the second approach, we take advantage of the fact that metabolic reactions show varying degrees of essentiality among different metabolic networks that are viable on the same carbon sources. Any one reaction can be essential in one such network and inessential in another, depending on which reactions and pathways are present in the network. One can quantify a reaction's degree of essentiality in randomly sampled viable networks via a "superessentiality index", defined as the fraction of metabolic networks in which the reaction is essential for viability on a given carbon source (22). Highly superessential reactions are essential in most random viable networks, and cannot be by-passed easily by alternative metabolic pathways. We first computed the superessentiality index of each reaction on each carbon source $C_i$, and assembled this information into a superessentiality vector. Each element of this vector corresponds to one of the 5,906 reactions in the reaction universe, and contains the fraction of random viable metabolic networks in which the reaction is essential for viability on $C_i$. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources $C_i$ and $C_j$ as a proxy for metabolic distance between the two carbon sources.

**S8: Distance measure between carbon sources based on superessential reactions**

In the second approach, we take advantage of the fact that metabolic reactions show varying degrees of essentiality among different metabolic networks that are viable on the same carbon sources. Any one reaction can be essential in one such network and inessential in another, depending on which reactions and pathways are present in the network. One can quantify a reaction's degree of essentiality in randomly sampled viable networks via a "superessentiality index", defined as the fraction of metabolic networks in which the reaction is essential for viability on a given carbon source (22). Highly superessential reactions are essential in most random viable networks, and cannot be by-passed easily by alternative metabolic pathways. We first computed the superessentiality index of each reaction on each carbon source $C_i$, and assembled this information into a superessentiality vector. Each element of this vector corresponds

to one of the 5,906 reactions in the reaction universe, and contains the fraction of random viable metabolic networks in which the reaction is essential for viability on $C_i$. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources $C_i$ and $C_j$ as a proxy for metabolic distance between the two carbon sources.

Previous work showed that highly superessential reactions are more likely to be involved in metabolic innovation (1). We thus also wanted to compute a biochemical distance measure of carbon sources based on this index. To this end, we computed, for each carbon source, the superessentiality index of all reactions belonging to the reaction universe, which yields a superessentiality vector of length 5,906. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources $C_i$ and $C_j$ as a proxy for the biochemical distance between the two carbon sources. Fig. S27a shows that the number of innovative offspring, which are generated by recombination between parents viable on $C_i$, and gain viability on a given carbon source $C_j$ is significantly correlated with the Euclidian distance between the superessentiality vectors for $(C_i, C_j)$ (Pearson $r$ = -0.3935, and $P < 10^{-83}$).

**S9: Random metabolic networks and erroneous energy generating cycles**

A recent study by Fritzemeier et al. showed that most of the published genome-scale metabolic networks include thermodynamically impossible energy-generating cycles (EGCs), which are capable of charging energy metabolites without nutrient consumption (25). It showed that these EGCs can artificially inflate biomass flux by 25% and could be particularly problematic in evolutionary simulations, which involves incorporation of foreign metabolic reactions from other species.

We applied the approach of Fritzemeier et al., to identify EGCs in metabolic networks (25), using 15 different energy dissipation reactions (EDRs) for each of the 15 different types of energy metabolites in the cell. (See https://doi.org/10.1371/journal.pcbi.1005494.s002 for complete information on these reactions). We maximized one energy dissipation reaction flux $v_d$ at a time, while preventing all influx of external nutrients into the model. The problem can be mathematically expressed as follows: $\max v_d$ subject to:

$$Sv = 0$$
$$\forall i \notin E: v_i^{min} \leq v_i \leq v_i^{max}$$
$$\forall i \in E: v_i = 0$$

where $S$ is the stoichiometric matrix describing a metabolic system, $v$ is the vector of all metabolic fluxes, $d$ is the index of one of the energy dissipation reactions, $v^{min}$ and $v^{max}$ are vectors of lower and upper reaction bounds, and $E$ is the set of indices of all exchange reactions. An optimal value $v_d^*$ for this optimization with $v_d^* > 0$ for at least one of the energy dissipation reactions demonstrates the existence of at least one EGC in the corresponding metabolic network.

Using this approach, we first determined that the initial *E. coli* metabolic network with 2079 reactions (12) from which we started most of our MCMC sampling had no EGCs. However, we found that 97.3% and 97.8% of our randomly sampled metabolic networks viable on glucose and acetate, respectively, harbored at least one EGC.

To determine whether these EGCs artificially inflated the number of innovative offspring, we sampled EGC-free parental metabolic networks. To do so, we modified our MCMC approach such that each sampled metabolic network not only retained viability in a given environment, but was also EGC-free. To fulfill these goals, we required that each step (reaction swap) in our MCMC sampling preserved viability on a given carbon source, and did not introduce an EGC (checked by the EGCs identification approach described above). Using this approach, we generated 1,000 pairs of EGC-free metabolic networks viable exclusively on glucose, and 1,000 pairs of EGC-free networks viable only on acetate. We then generated 1,000 recombinant offspring from each pair. Recombination between EGC-free metabolisms viable exclusively on glucose resulted in 29,941 innovative offspring, only 7.41% fewer than the corresponding number for EGC-containing metabolisms (32,338). Likewise, we observed 46,941 innovative offspring emerging from EGC-free parental metabolisms viable exclusively on acetate, 5.57% fewer than the corresponding number for EGC-containing metabolisms (49,708). Thus, removing EGCs slightly reduces the incidence of innovation (figure S30). Importantly, the patterns of relative constraints remain almost exactly unchanged (figure S31).

Fritzemeier et al. showed that EGCs could artificially increase the biomass rate of metabolic networks by 25% (25). However, figure S32 indicates that the majority of viable networks we study already have a biomass flux considerably larger than our threshold of viability, so reducing their biomass production rate by 25% will not result in a viability loss for most metabolisms, which is why excluding EGCs does not

substantially reduce the emergence of novel phenotypes.

**Supplementary References:**


1. Hosseini, S.-R., O. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc. R. Soc. B. .
2. Edwards, J.S., R.U. Ibarra, and B.O. Palsson. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19: 125–30.
3. Edwards, J.S., and B.O. Palsson. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274: 17410–6.
4. Lewis, N.E., H. Nagarajan, and B.O. Palsson. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. 10: 291–305.
5. Feist, A.M., and B.Ø. Palsson. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat. Biotechnol. 26: 659–67.
6. McCloskey, D., B.Ø. Palsson, and A.M. Feist. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol. 9: 661.
7. Goto, S., T. Nishioka, and M. Kanehisa. 2000. LIGAND: chemical database of enzyme reactions. Nucleic Acids Res. 28: 380–2.
8. Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 30: 402–4.
9. Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38: D355-60.
10. Matias Rodrigues, J.F., and A. Wagner. 2009. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput. Biol. 5: e1000613.
11. Orth, J.D., I. Thiele, and B.Ø. Palsson. 2010. What is flux balance analysis? Nat. Biotechnol. 28: 245–8.
12. Feist, A.M., C.S. Henry, J.L. Reed, M. Krummenacker, A.R. Joyce, P.D. Karp, L.J. Broadbelt, V. Hatzimanikatis, and B.Ø. Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3: 121.
13. King, Z.A., J. Lu, A. Dräger, P. Miller, S. Federowicz, J.A. Lerman, A. Ebrahim, B.O. Palsson, and N.E. Lewis. 2015. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. : gkv1049-.
14. Lercher, M.J., and C. Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25: 559–67.

15. Ibarra, R.U., J.S. Edwards, and B.O. Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 420: 186–9.

16. Vieira-Silva, S., and E.P.C. Rocha. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 6: e1000808.

17. Kirschner, D., and S. Marino. 2005. Mycobacterium tuberculosis as viewed through a computer. Trends Microbiol. 13: 206–11.

18. Fong, S.S., and B.Ø. Palsson. 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat. Genet. 36: 1056–8.

19. Fong, S.S., J.Y. Marciniak, and B.O. Palsson. 2003. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J. Bacteriol. 185: 6400–6408.

20. Samal, A., J.F. Matias Rodrigues, J. Jost, O.C. Martin, and A. Wagner. 2010. Genotype networks in metabolic reaction spaces. BMC Syst. Biol. 4: 30.

21. Wagner, A., and D.A. Fell. 2001. The small world inside large metabolic networks. Proc. R. Soc. B Biol. Sci. 268: 1803–1810.

22. Barve, A., J.F.M. Rodrigues, and A. Wagner. 2012. Superessential reactions in metabolic networks. Proc. Natl. Acad. Sci. U. S. A. 109: E1121-30.

23. Ma, H.-W., and A.-P. Zeng. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19: 1423–30.

24. Hopcroft, J., and R. Tarjan. 1973. Algorithm 447: efficient algorithms for graph manipulation. Commun. ACM. 16: 372–378.

25. Fritzemeier, C.J., D. Hartleb, B. Szappanos, B. Papp, M.J. Lercher, and G. Fekete. 2017. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Comput. Biol. 13: e1005494.

## Supplementary Figures

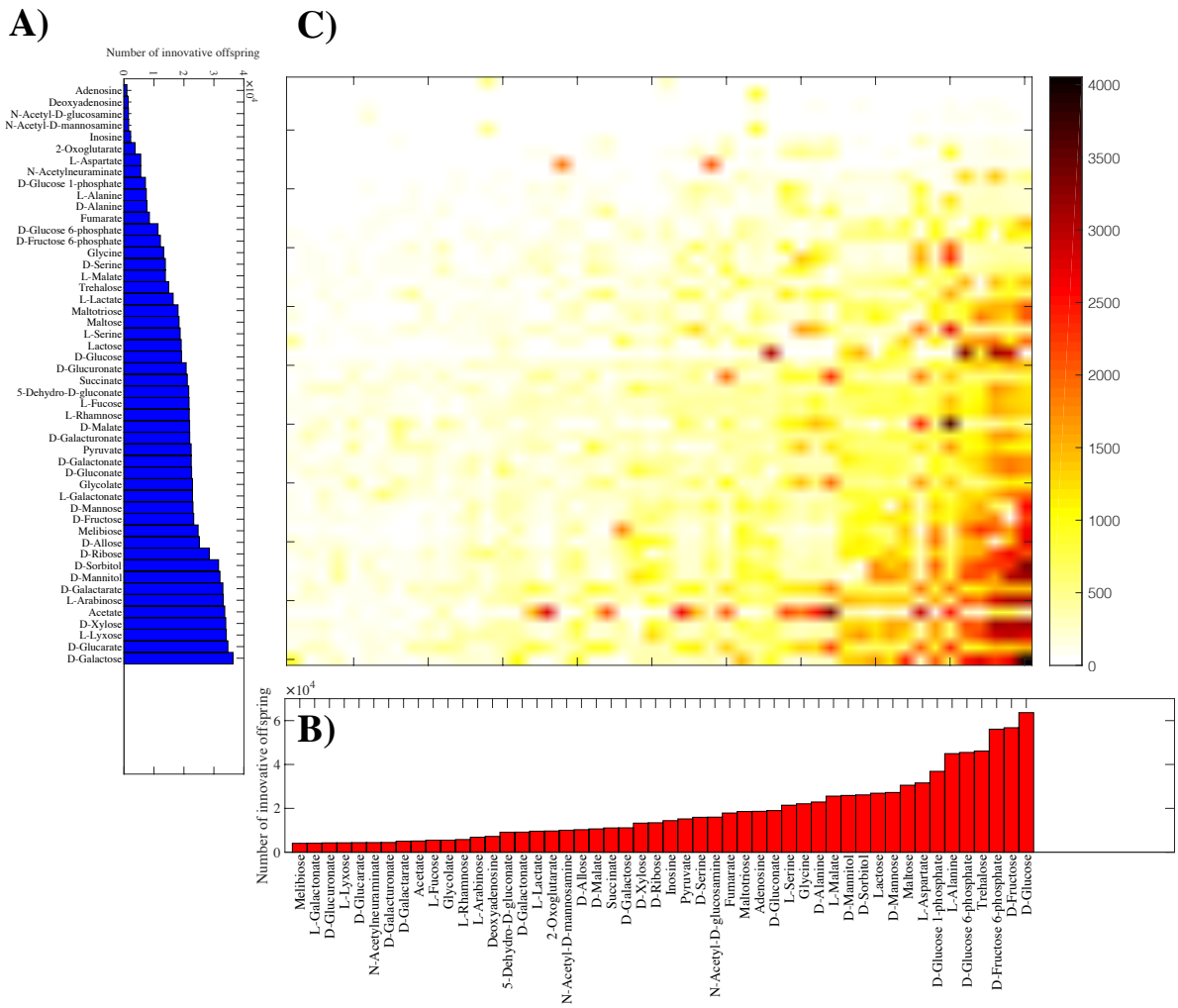



**Figure S1: Recombination can create all 50 carbon-use phenotypes considered here ($n$ = 20). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 37, ranging from 977 on Adenosine to 356,378 on D-galactose. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the $x$-axis. This number varies by a factor 15, ranging from 4,042 on melibiose to 63,634 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 20 reactions are swapped between parental metabolic networks in a recombination event.

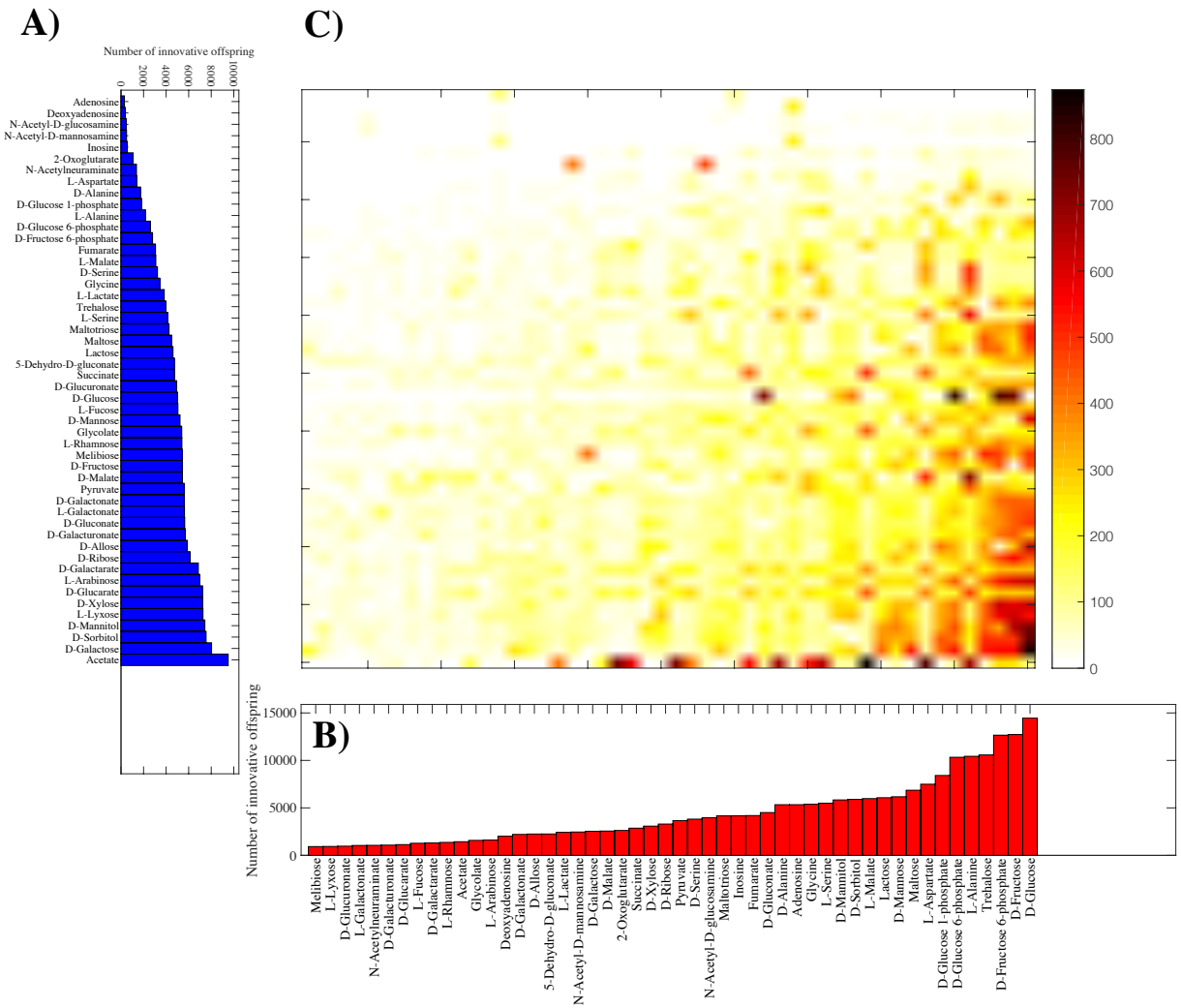



**Figure S2: Recombination can create all 50 carbon-use phenotypes considered here ($n$ = 30). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 32, ranging from 299 on adenosine to 9,503 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the $x$-axis. This number varies by a factor 16, ranging from 923 on melibiose to 14,452 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 30 reactions are swapped between parental metabolic networks in a recombination event.

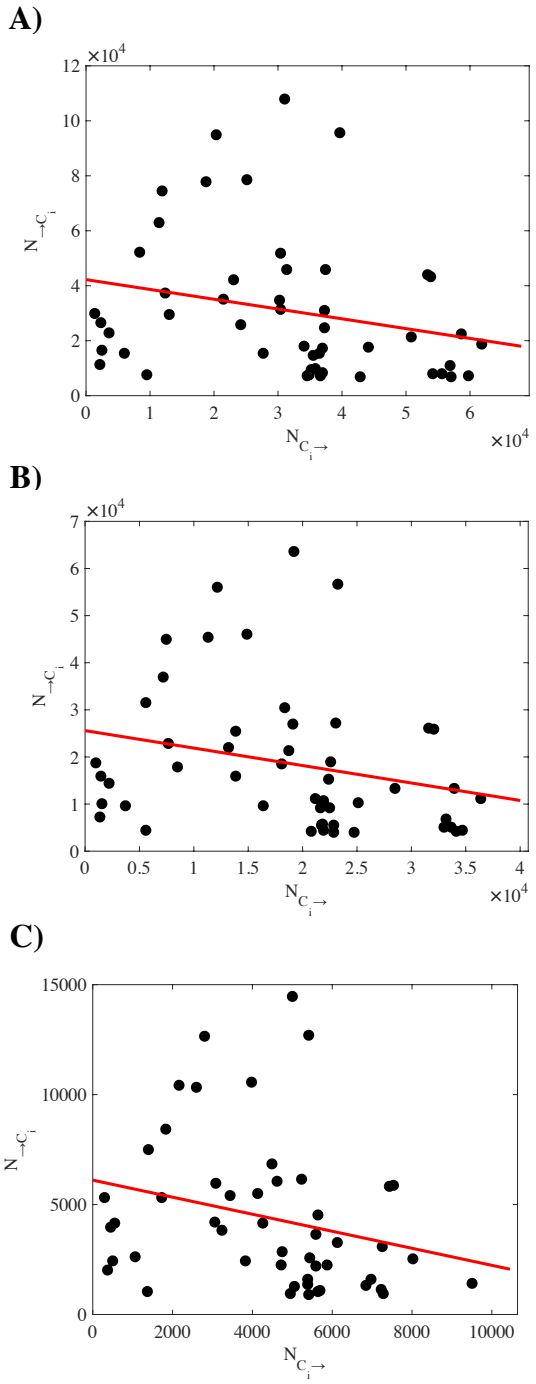



**Figure S3: Negative correlation between** ($N_{C_i\rightarrow}$) **and** ($N_{\rightarrow C_i}$)**.** Each circle corresponds to a given carbon source $C_i$. The vertical axis shows ($N_{C_i\rightarrow}$), the number of metabolic innovations emerging from parents viable on carbon source $C_i$. The horizontal axis shows ($N_{\rightarrow C_i}$), the number of innovations leading to viability on $C_i$. There is a negative correlation between ($N_{C_i\rightarrow}$) and ($N_{\rightarrow C_i}$), regardless of the number ($n$) of reactions exchanged: **A)** ($n$ = 10, Pearson $r$ = -0.239, P < 0.093), **B)** ($n$ = 20, Pearson $r$ = -0.248, P < 0.082), **C)** ($n$ = 30, Pearson $r$ = -0.256, P < 0.073). For all analyses the genotypic distance between parents is $D$ = 100.

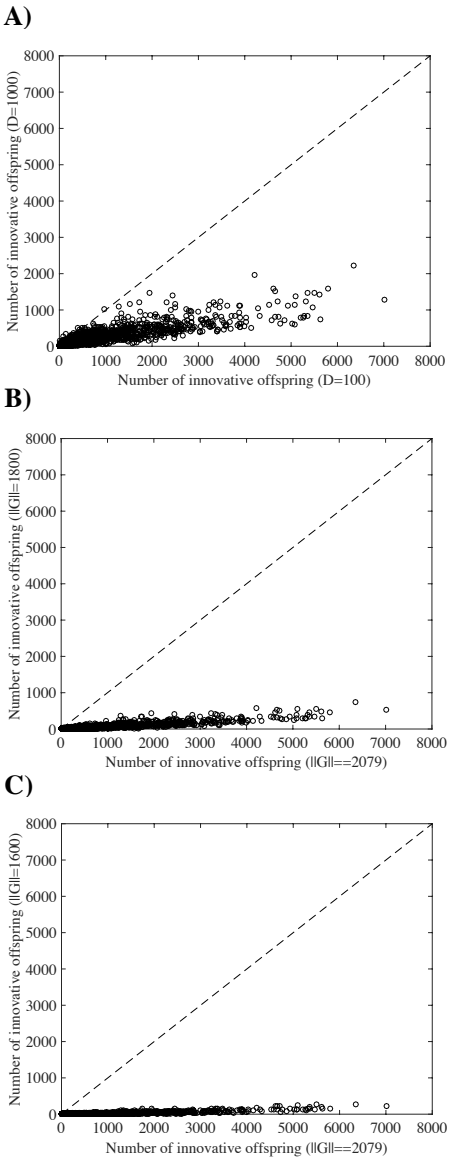



**Figure S4: Fewer innovative offspring at higher genotypic distance (*D*) and smaller metabolic network size ||*G*||.** Each circle corresponds to a pair of carbon sources $(C_i, C_j)$ and shows the number of innovative offspring gaining viability on $C_j$, which are generated by recombination between parents viable on carbon source $C_i$. The horizontal axis specifies the number of innovative offspring where parents have genetic distance $D$ = 100, and metabolic network size $||G||$ = 2,079. The vertical axes provide the same information, but for parents with A) genotypic distance $D$ = 1,000, and metabolic network size $||G||$ = 2,079 reactions, B) genotypic distance $D$ = 100, and metabolic network size $||G||$ = 1,800 reactions, and C) genotypic distance $D$ = 100, and metabolic network size $||G||$ = 1,600 reactions. The dashed diagonal lines correspond to the identity line ($y = x$). Note that in all three panels, most or all data lie below this line, indicating that higher parental genotypic distance and lower metabolic network size lead to fewer innovative offspring for almost all carbon source pair.

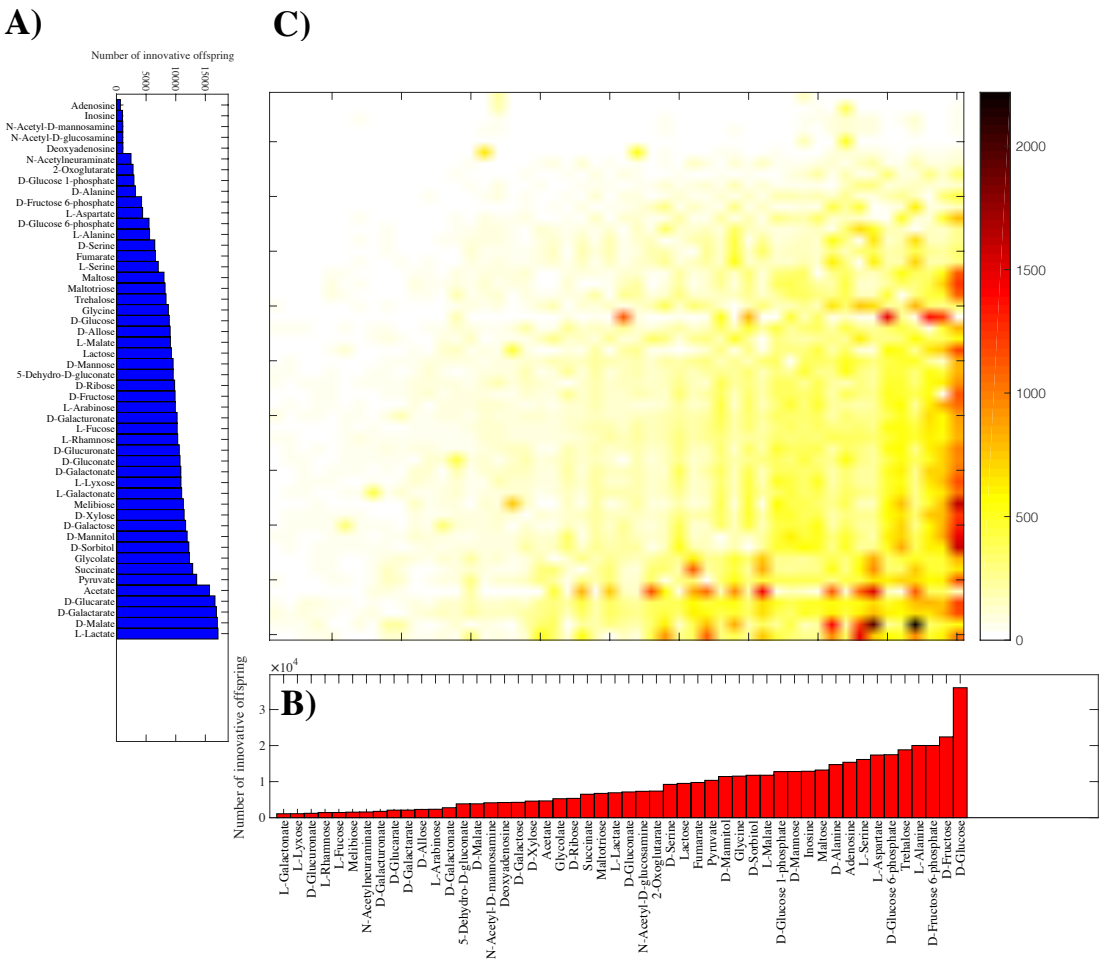



**Figure S5: Recombination can create all 50 carbon-use phenotypes considered here (*D* = 1,000). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 25, ranging from 662 on adenosine to 17,132 on L-lactate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 33, ranging from 1081 on L-galactonate to 36,051 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same number as the *E. coli* metabolic network, and they differ in $D = 1{,}000$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

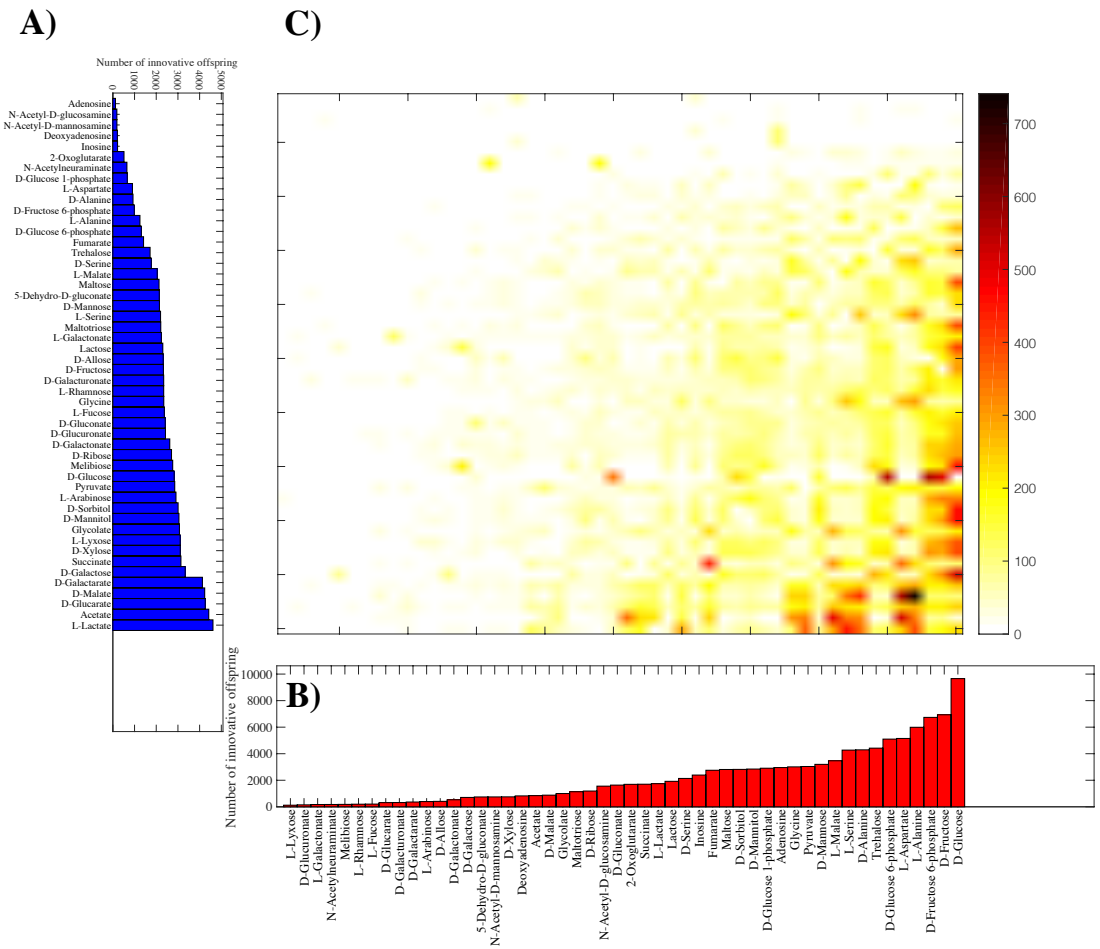



**Figure S6: Recombination can create all 50 carbon-use phenotypes considered here (||*G*|| = 1800). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 38, ranging from 120 on adenosine to 4,616 on L-lactate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 79, ranging from 122 on L-lyxose to 9,657 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain ||*G*|| = 1,800 reactions and differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

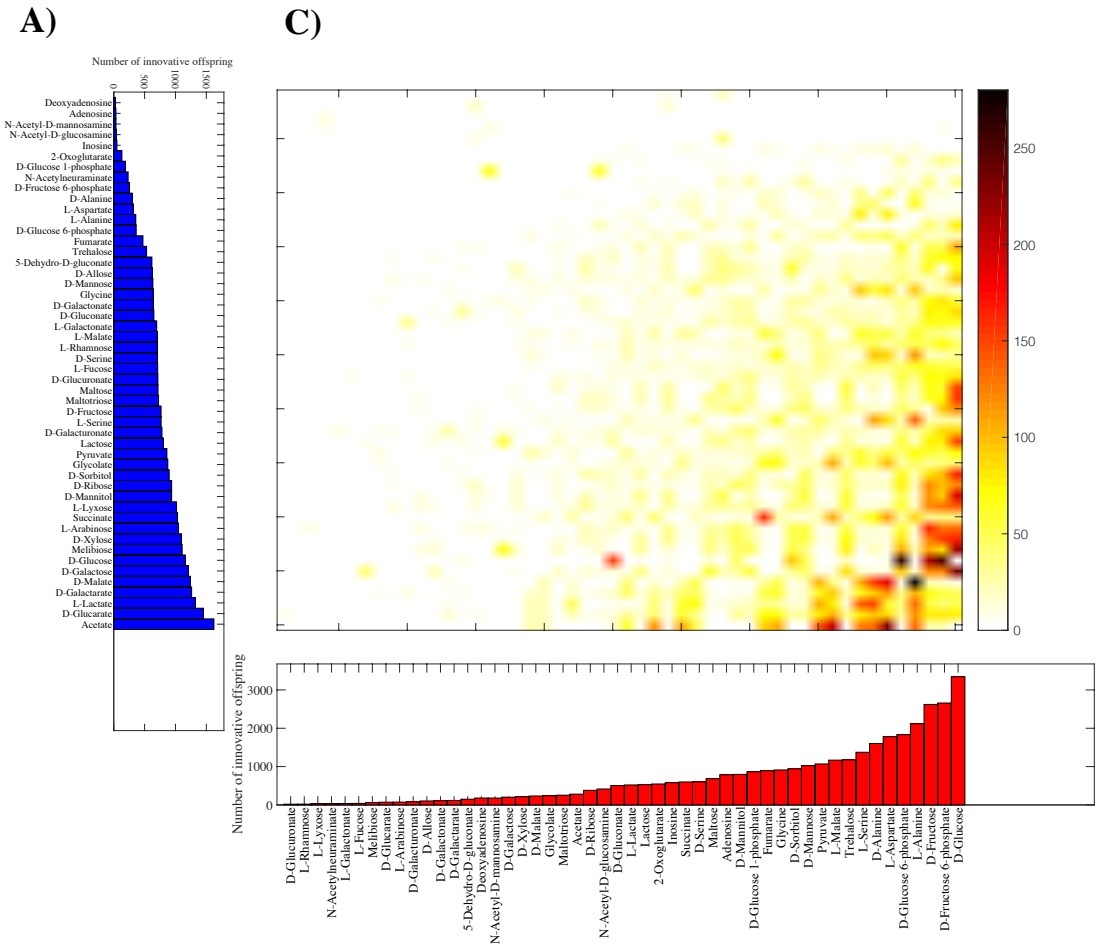



**Figure S7: Recombination can create all 50 carbon-use phenotypes considered here (||*G*|| = 1,600). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 58, ranging from 28 on deoxyadenosine to 1,623 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 176, ranging from 19 on D-glucuronate to 3,344 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain ||*G*|| = 1,600 reactions and differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

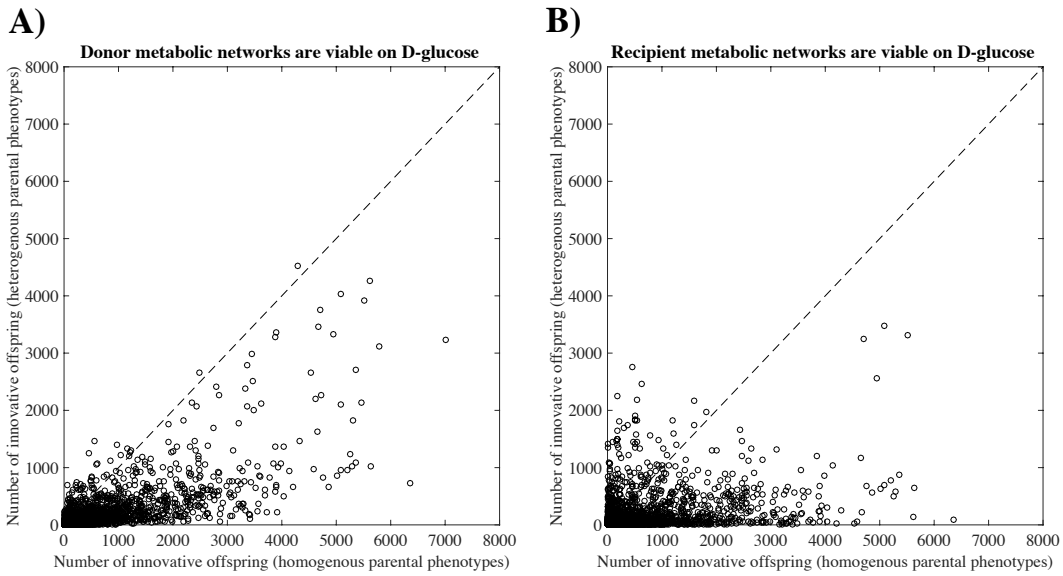



**Figure S8: Fewer innovative offspring from phenotypically heterogeneous parents than from phenotypically homogenous parents.** Each circle corresponds to a given pair of carbon sources $(C_i, C_j)$ and shows the number of innovative offspring gaining viability on $C_j$, that are generated by recombination between parents viable on carbon source $C_i$. The horizontal axis specifies the number of innovative offspring for parents that are viable on the same carbon sources (phenotypically homogeneous parents). The vertical axes show the number of innovative offspring for **A)** parental donors viable on D-glucose and parental recipients viable on $C_i$, and **B)** parental recipients are viable on D-glucose, and parental donors viable on $C_i$. In these analyses, all parents have $||G||$ = 2,079 reactions, the same as the *E. coli* metabolic network, and their genotypic distance ($D$) is constant and equals 100. Note that in both panels, the majority of circles (with few exceptions) are placed below the identity ($y = x$) line, indicating that it is more likely for phenotypically homogenous parents to generate innovative offspring than for phenotypically heterogeneous parents.

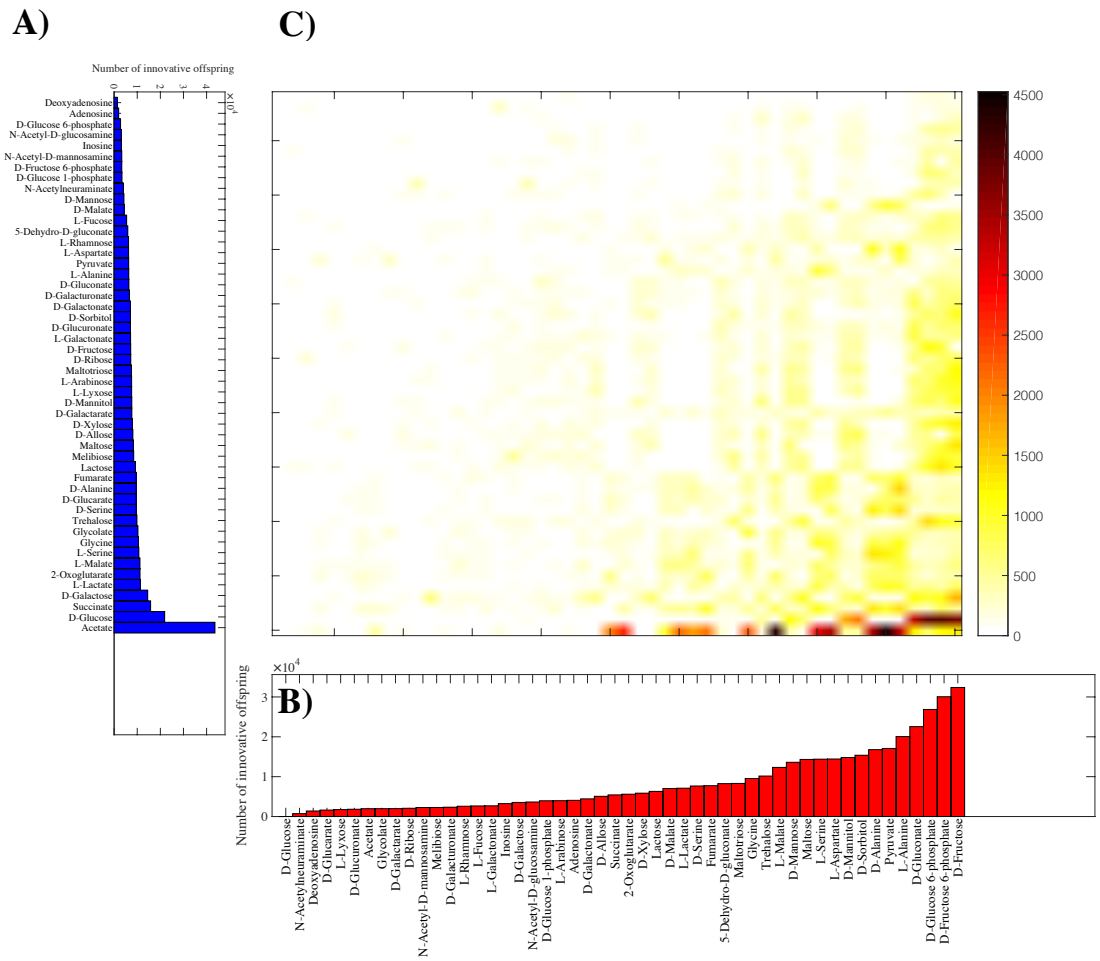



**Figure S9: Recombination can create all 50 carbon-use phenotypes considered here (Parents with heterogeneous phenotypes, donors viable only on glucose). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between donor parents viable on glucose and recipient parents that are viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 32, ranging from 1,371 on deoxyadenosine to 43,615 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 44, ranging from 729 on N-acetylneuraminate to 32,378 on D-fructose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between donor parents viable on glucose, and recipient parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same number as in the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

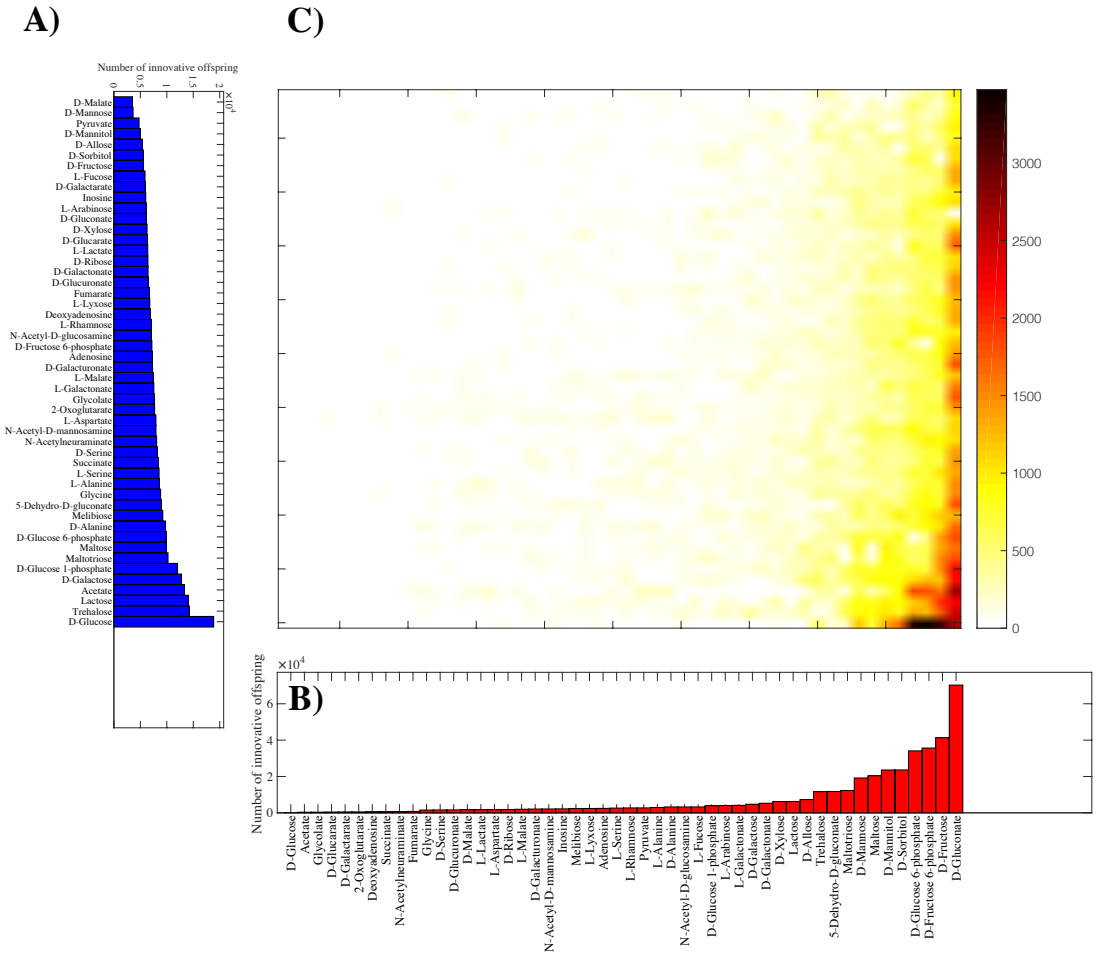
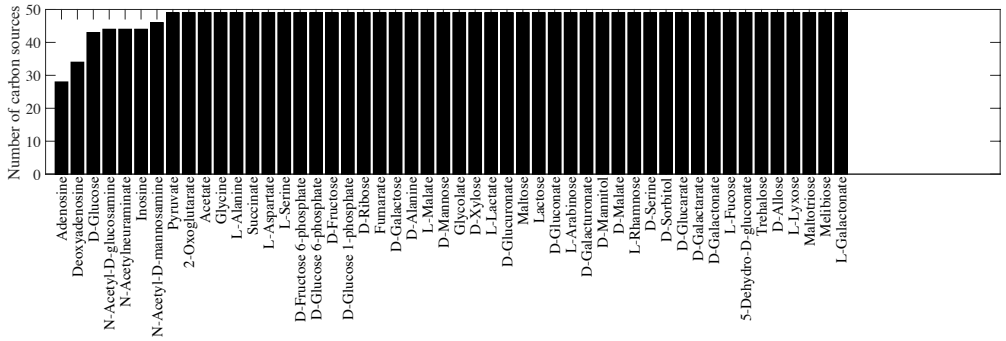
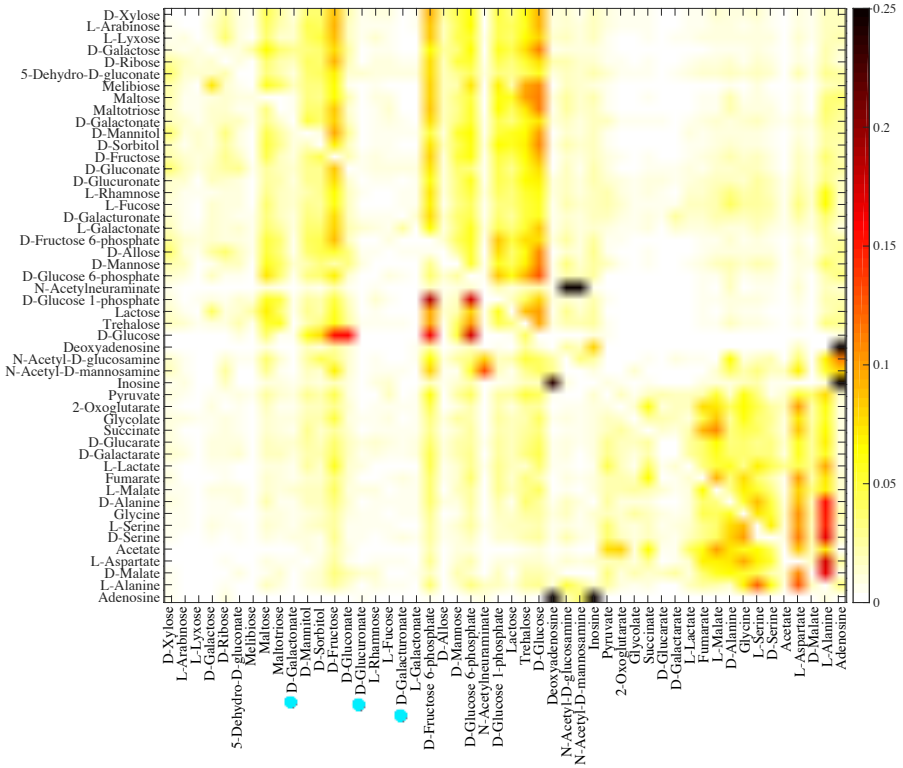



**Figure S10: Recombination can create all 50 carbon-use phenotypes considered here (Parents with heterogeneous phenotypes, recipients viable only on glucose). A)** The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between recipient parents viable on glucose and donor parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 5, ranging from 3,511 on D-malate to 18,856 on D-glucose. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 204, ranging from 343 on acetate to 70,292 on D-gluconate. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between recipient parents viable on glucose, and donor parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same as in the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Xylose
L-Arabinose
L-Lyxose
D-Galactose
D-Ribose
5-Dehydro-D-gluconate
Melibiose
Maltose
Maltotriose
D-Galactonate
D-Mannitol
D-Sorbitol
D-Fructose
D-Gluconate
D-Glucuronate
L-Rhamnose
L-Fucose
D-Galacturonate
L-Galactonate
D-Fructose 6-phosphate
D-Allose
D-Mannose
D-Glucose 6-phosphate
N-Acetylneuraminate
D-Glucose 1-phosphate
Lactose
Trehalose
D-Glucose
Deoxyadenosine
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Inosine
Pyruvate
2-Oxoglutarate
Glycolate
Succinate
D-Glucarate
D-Galactarate
L-Lactate
Fumarate
L-Malate
D-Alanine
Glycine
L-Serine
D-Serine
Acetate
L-Aspartate
D-Malate
L-Alanine
Adenosine
0.25
0.2
0.15
0.1
0.05
0

**C)**

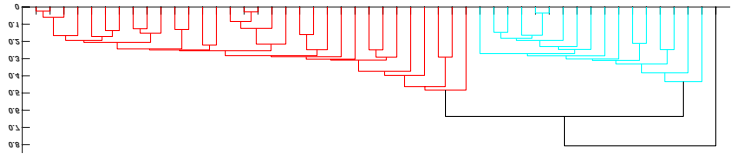
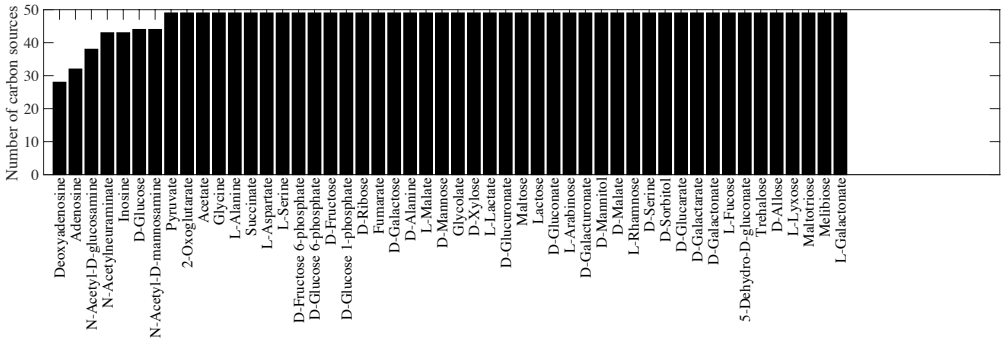
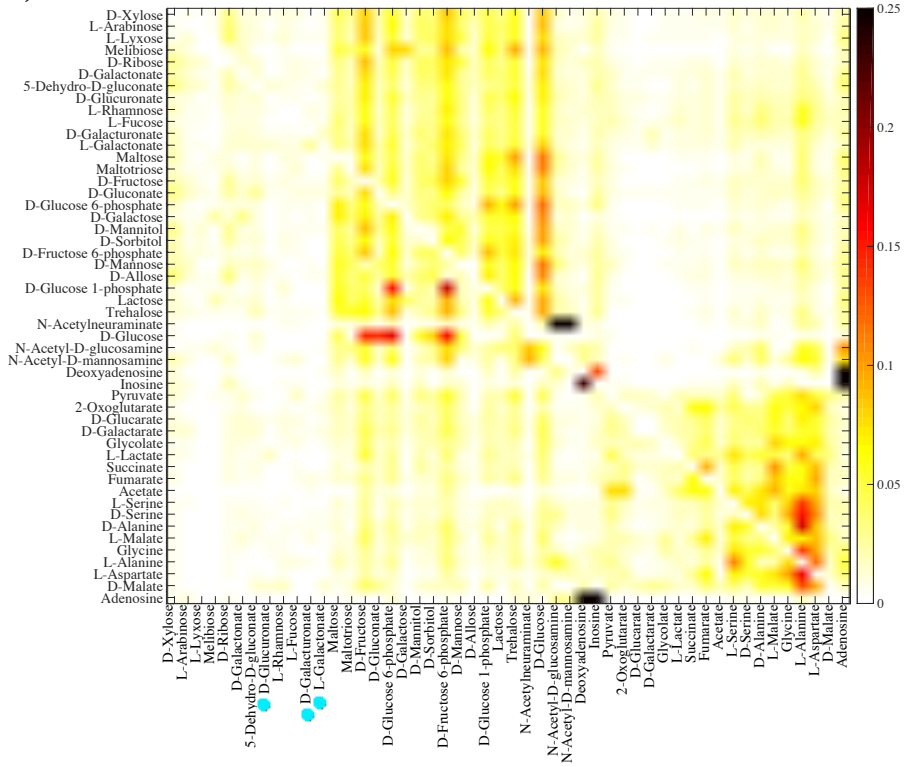


**Figure S11: Emergence of innovative offspring can be constrained by parental phenotypes ($n$ = 20). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate, (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 20 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
D-Glucose
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Xylose
L-Arabinose
L-Lyxose
Melibiose
D-Ribose
D-Galactonate
5-Dehydro-D-gluconate
D-Glucuronate
L-Rhamnose
L-Fucose
D-Galacturonate
L-Galactonate
Maltose
Maltotriose
D-Fructose
D-Gluconate
D-Glucose 6-phosphate
D-Galactose
D-Mannitol
D-Sorbitol
D-Fructose 6-phosphate
D-Mannose
D-Allose
D-Glucose 1-phosphate
Lactose
Trehalose
N-Acetylneuraminate
D-Glucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
Pyruvate
2-Oxoglutarate
D-Glucarate
D-Galactarate
Glycolate
L-Lactate
Succinate
Fumarate
Acetate
L-Serine
D-Serine
D-Alanine
L-Malate
Glycine
L-Alanine
L-Aspartate
D-Malate
Adenosine
0.25
0.2
0.15
0.1
0.05
0

**C)**

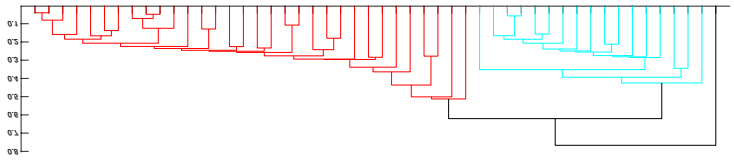
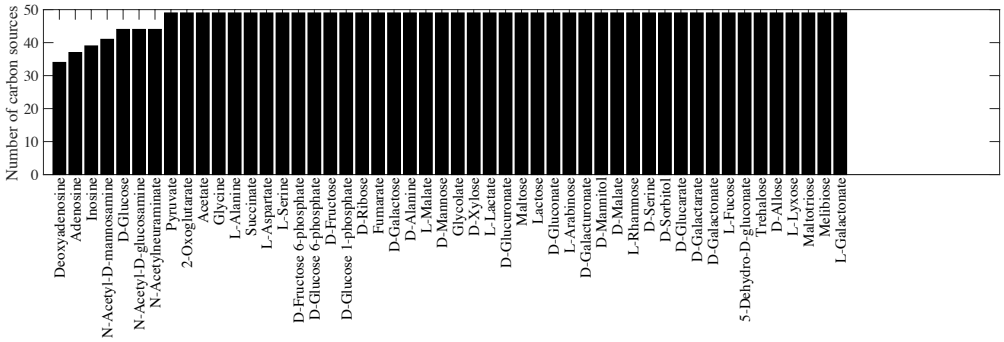
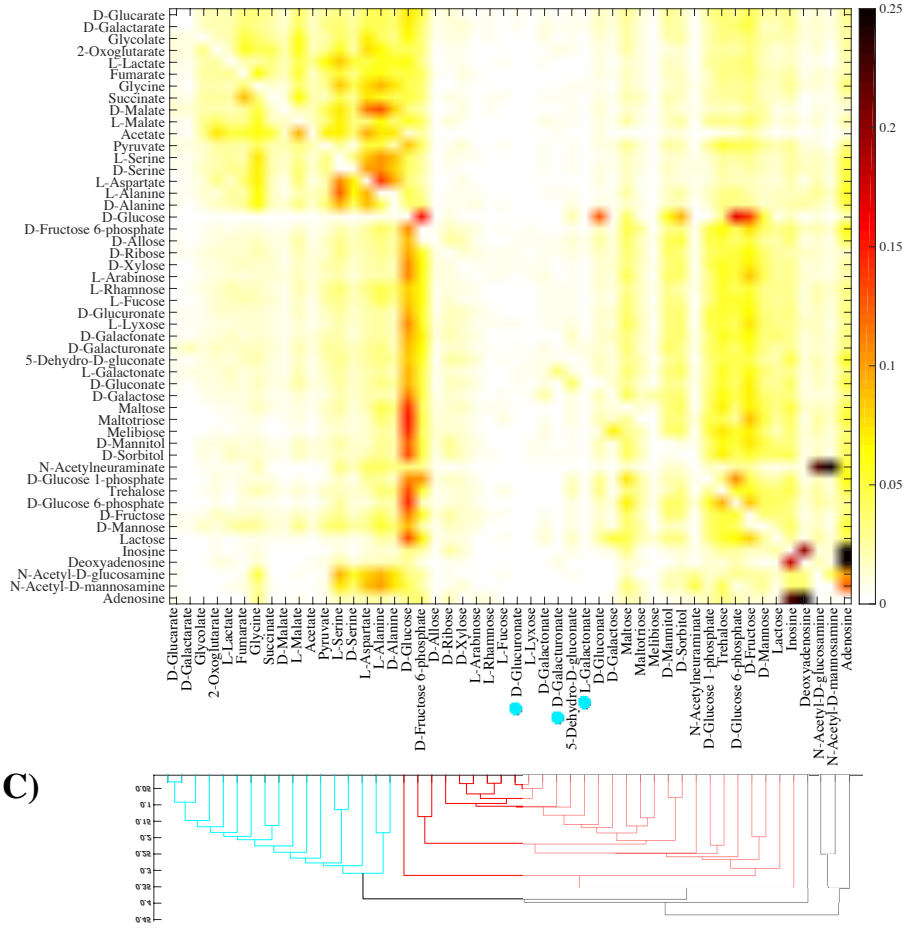


**Figure S12: Emergence of innovative offspring can be constrained by parental phenotypes ($n$ = 30). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 30 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

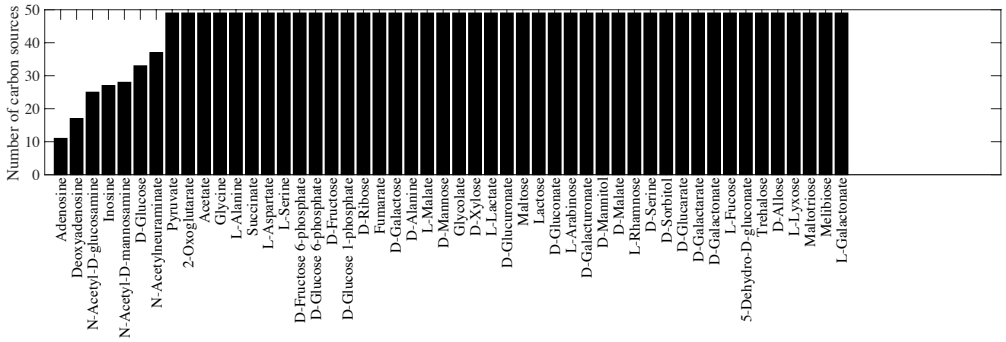
**B)**


D-Glucarate
D-Galactarate
Glycolate
2-Oxoglutarate
L-Lactate
Fumarate
Glycine
Succinate
D-Malate
L-Malate
Acetate
Pyruvate
L-Serine
D-Serine
L-Aspartate
L-Alanine
D-Alanine
D-Glucose
D-Fructose 6-phosphate
D-Allose
D-Ribose
D-Xylose
L-Arabinose
L-Rhamnose
L-Fucose
D-Glucuronate
L-Lyxose
D-Galactonate
D-Galacturonate
5-Dehydro-D-gluconate
L-Galactonate
D-Gluconate
D-Galactose
Maltose
Maltotriose
Melibiose
D-Mannitol
D-Sorbitol
N-Acetylneuraminate
D-Glucose 1-phosphate
Trehalose
D-Glucose 6-phosphate
D-Fructose
D-Mannose
Lactose
Inosine
Deoxyadenosine
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
0
0.05
0.1
0.15
0.2
0.25

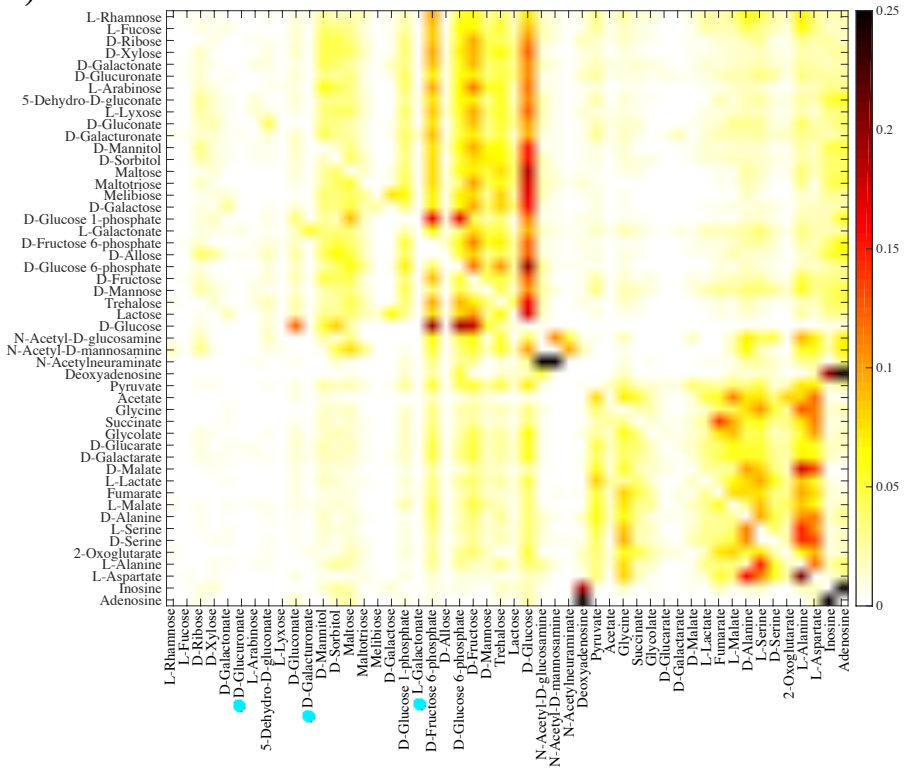
**C)**

**Figure S13: Emergence of innovative offspring can be constrained by parental phenotypes (*D* = 1,000). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 1,000 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


L-Rhamnose
L-Fucose
D-Ribose
D-Xylose
D-Galactonate
D-Glucuronate
L-Arabinose
5-Dehydro-D-gluconate
L-Lyxose
D-Gluconate
D-Galacturonate
D-Mannitol
D-Sorbitol
Maltose
Maltotriose
Melibiose
D-Galactose
D-Glucose 1-phosphate
L-Galactonate
D-Fructose 6-phosphate
D-Allose
D-Glucose 6-phosphate
D-Fructose
D-Mannose
Trehalose
Lactose
D-Glucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
N-Acetylneuraminate
Deoxyadenosine
Pyruvate
Acetate
Glycine
Succinate
Glycolate
D-Glucarate
D-Galactarate
D-Malate
L-Lactate
Fumarate
L-Malate
D-Alanine
L-Serine
D-Serine
2-Oxoglutarate
L-Alanine
L-Aspartate
Inosine
Adenosine
0.25
0.2
0.15
0.1
0.05
0

**C)**

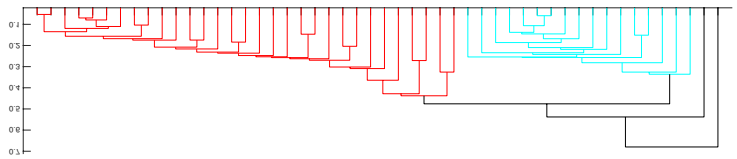
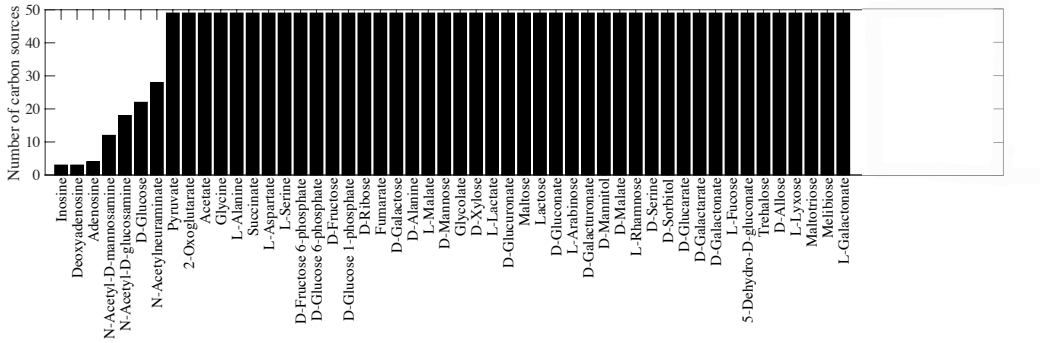
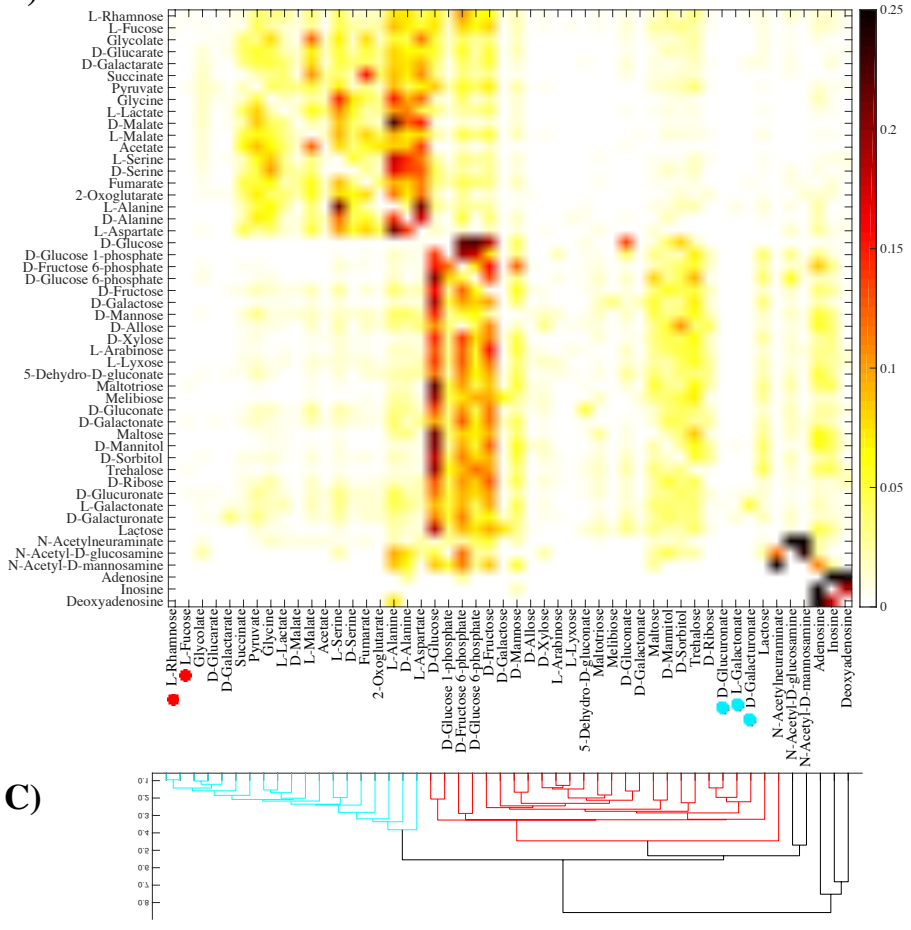

0.4
0.3
0.3
0.4
0.2
0.0
0.1

**Figure S14: Emergence of innovative offspring can be constrained by parental phenotypes (||*G*|| = 1,800). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain ||*G*|| = 1,800 reactions, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Inosine
Deoxyadenosine
Adenosine
N-Acetyl-D-mannosamine
N-Acetyl-D-glucosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

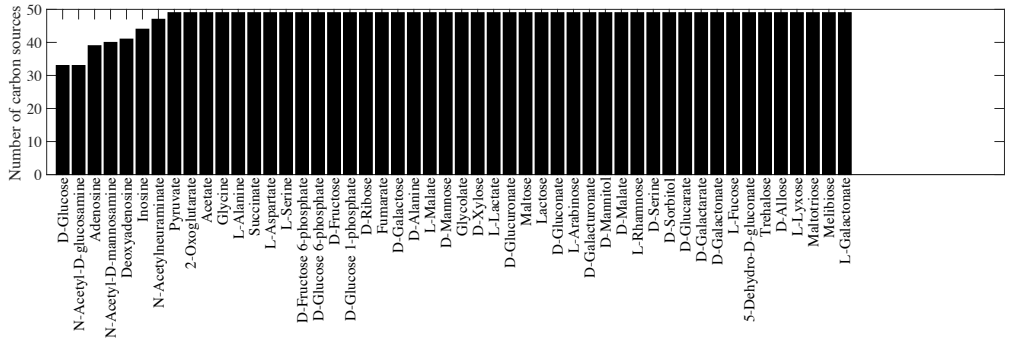
**B)**


L-Rhamnose
L-Fucose
Glycolate
D-Glucarate
D-Galactarate
Succinate
Pyruvate
Glycine
L-Lactate
D-Malate
L-Malate
Acetate
L-Serine
D-Serine
Fumarate
2-Oxoglutarate
L-Alanine
D-Alanine
L-Aspartate
D-Glucose
D-Glucose 1-phosphate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Galactose
D-Mannose
D-Allose
D-Xylose
L-Arabinose
L-Lyxose
5-Dehydro-D-gluconate
Maltotriose
Melibiose
D-Gluconate
D-Galactonate
Maltose
D-Mannitol
D-Sorbitol
Trehalose
D-Ribose
D-Glucuronate
L-Galactonate
D-Galacturonate
Lactose
N-Acetylneuraminate
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
Inosine
Deoxyadenosine
0.25
0.2
0.15
0.1
0.05
0

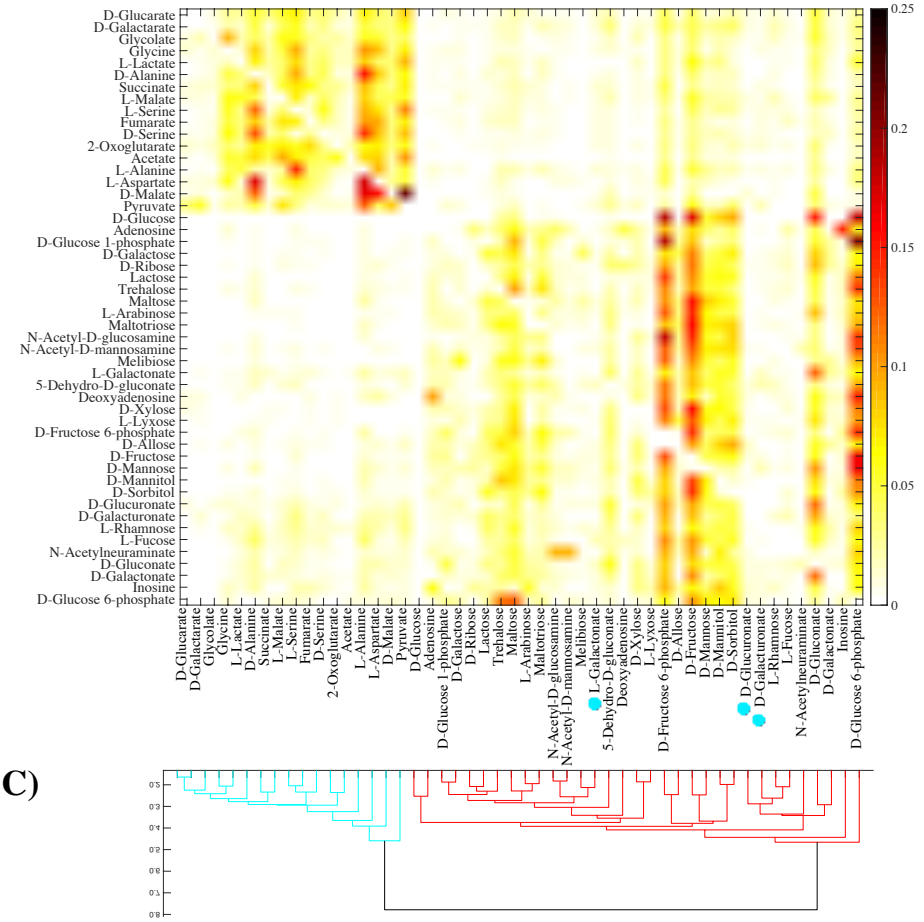
**C)**

**Figure S15: Emergence of innovative offspring can be constrained by parental phenotypes (||*G*|| = 1,600). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). In these analyses, parental metabolic networks contain ||*G*|| = 1,600 reactions, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
D-Glucose
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**

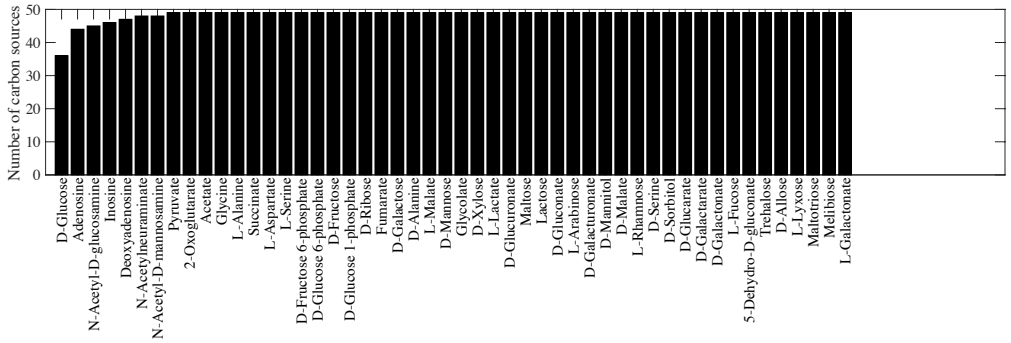

D-Glucarate
D-Galactarate
Glycolate
Glycine
L-Lactate
D-Alanine
Succinate
L-Malate
L-Serine
Fumarate
D-Serine
2-Oxoglutarate
Acetate
L-Alanine
L-Aspartate
D-Malate
Pyruvate
D-Glucose
Adenosine
D-Glucose 1-phosphate
D-Galactose
D-Ribose
Lactose
Trehalose
Maltose
L-Arabinose
Maltotriose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Melibiose
L-Galactonate
5-Dehydro-D-gluconate
Deoxyadenosine
D-Xylose
L-Lyxose
D-Fructose 6-phosphate
D-Allose
D-Fructose
D-Mannose
D-Mannitol
D-Sorbitol
D-Glucuronate
D-Galacturonate
L-Rhamnose
L-Fucose
N-Acetylneuraminate
D-Gluconate
D-Galactonate
Inosine
D-Glucose 6-phosphate
0
0.05
0.1
0.15
0.2
0.25

**C)**

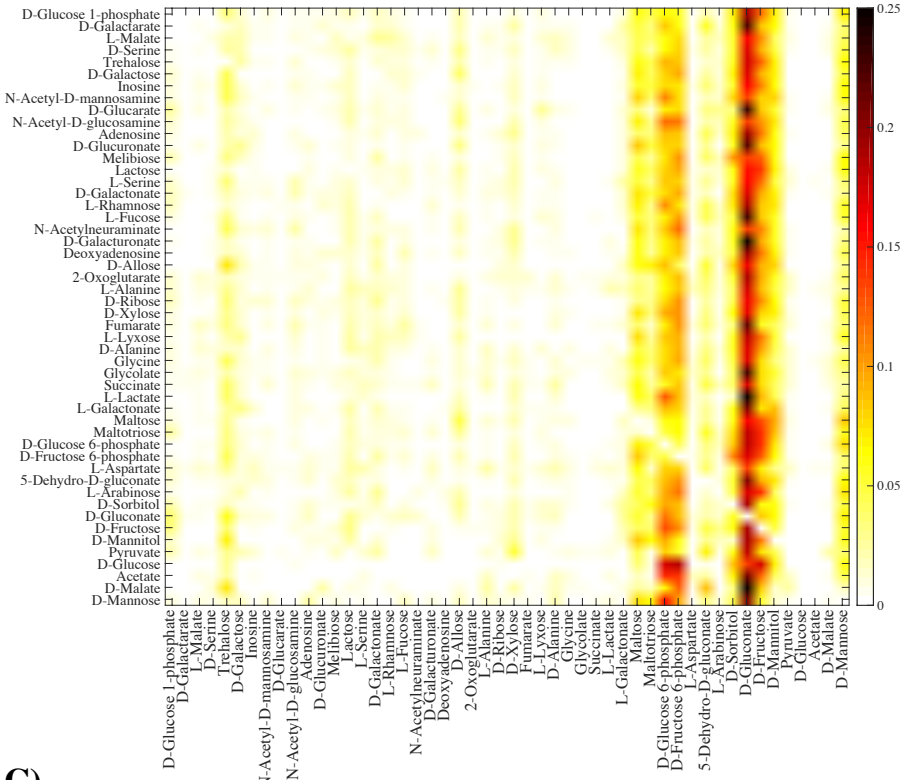
**Figure S16: Emergence of innovative offspring can be constrained by parental phenotypes (Parents with heterogeneous phenotypes, donors viable only on glucose ). A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between donor parents viable on glucose and the recipient parents viable exclusively on the carbon source specified on the vertical axis., which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same number as in the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

=

**A)**


Number of carbon sources
0
10
20
30
40
50
D-Glucose
Adenosine
N-Acetyl-D-glucosamine
Inosine
Deoxyadenosine
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Glucose 1-phosphate
D-Galactarate
L-Malate
D-Serine
Trehalose
D-Galactose
Inosine
N-Acetyl-D-mannosamine
D-Glucarate
N-Acetyl-D-glucosamine
Adenosine
D-Glucuronate
Melibiose
Lactose
L-Serine
D-Galactonate
L-Rhamnose
L-Fucose
N-Acetylneuraminate
D-Galacturonate
Deoxyadenosine
D-Allose
2-Oxoglutarate
L-Alanine
D-Ribose
D-Xylose
Fumarate
L-Lyxose
D-Alanine
Glycine
Glycolate
Succinate
L-Lactate
L-Galactonate
Maltose
Maltotriose
D-Glucose 6-phosphate
D-Fructose 6-phosphate
L-Aspartate
5-Dehydro-D-gluconate
L-Arabinose
D-Sorbitol
D-Gluconate
D-Fructose
D-Mannitol
Pyruvate
D-Glucose
Acetate
D-Malate
D-Mannose
0.25
0.2
0.15
0.1
0.05
0

**C)**

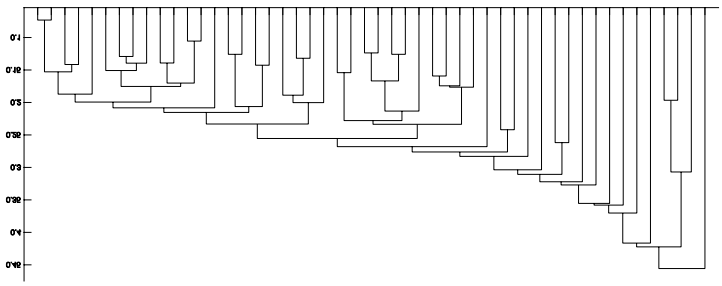
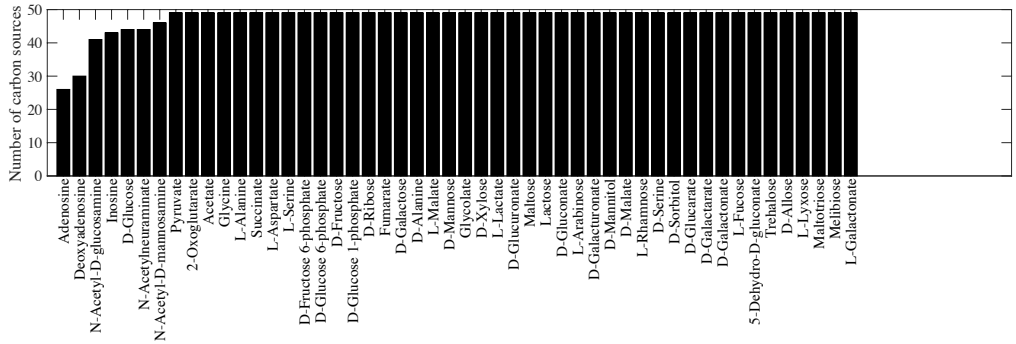
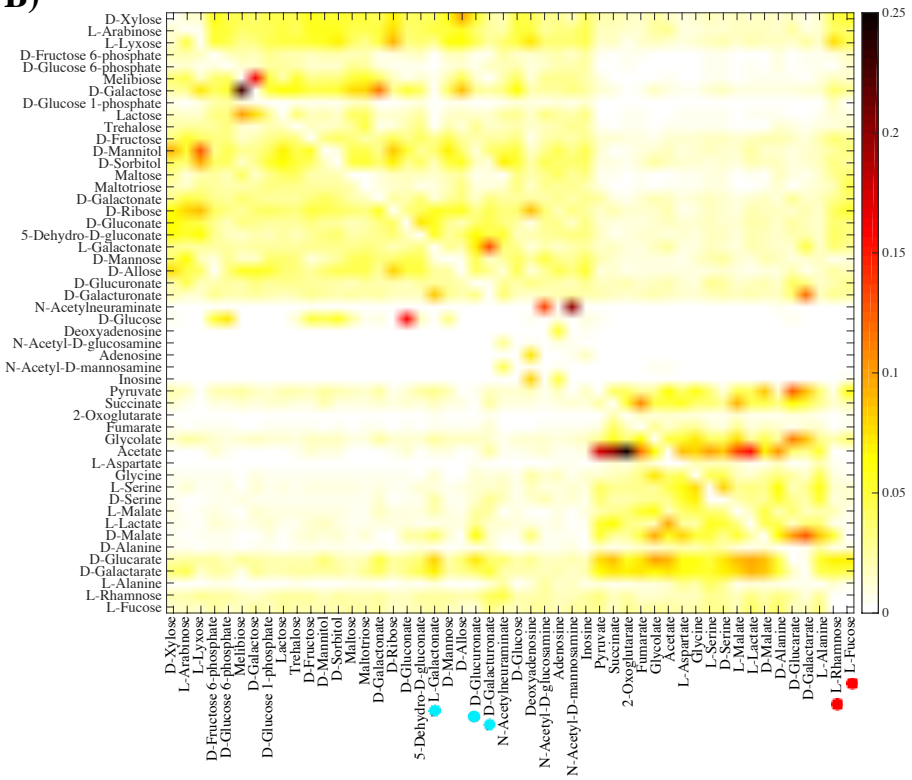


**Figure S17: Emergence of innovative offspring can be constrained by parental phenotypes (Parents with heterogeneous phenotypes, recipients viable only on glucose ).** **A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between recipient parents viable on glucose and donor parents viable exclusively on the carbon source specified on the vertical axis., which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. In this figure, main branches do not reflect glycolytic and gluconeogenic carbon sources as in other figures. In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same number as in the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
D-Glucose
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Xylose
L-Arabinose
L-Lyxose
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Melibiose
D-Galactose
D-Glucose 1-phosphate
Lactose
Trehalose
D-Fructose
D-Mannitol
D-Sorbitol
Maltose
Maltotriose
D-Galactonate
D-Ribose
D-Gluconate
5-Dehydro-D-gluconate
L-Galactonate
D-Mannose
D-Allose
D-Glucuronate
D-Galacturonate
N-Acetylneuraminate
D-Glucose
Deoxyadenosine
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Inosine
Pyruvate
Succinate
2-Oxoglutarate
Fumarate
Glycolate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Malate
L-Lactate
D-Malate
D-Alanine
D-Glucarate
D-Galactarate
L-Alanine
L-Rhamnose
L-Fucose
0
0.05
0.1
0.15
0.2
0.25

**C)**

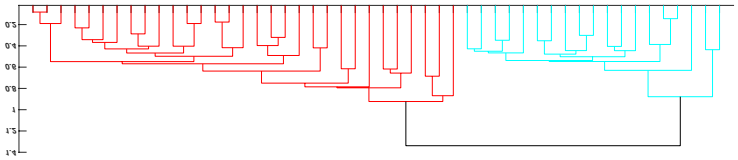
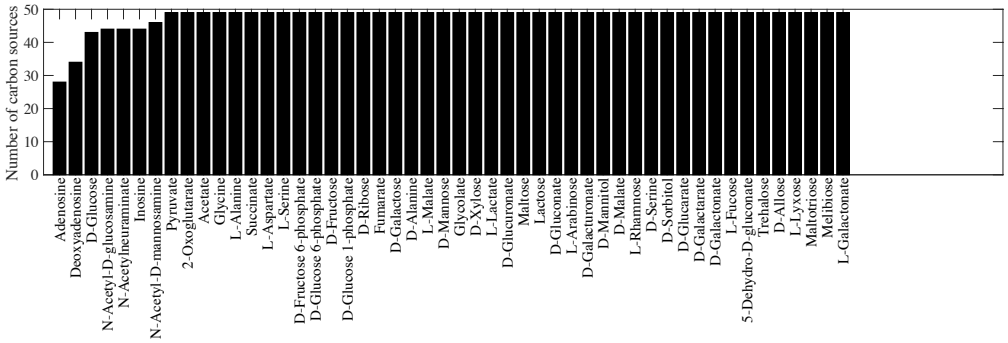
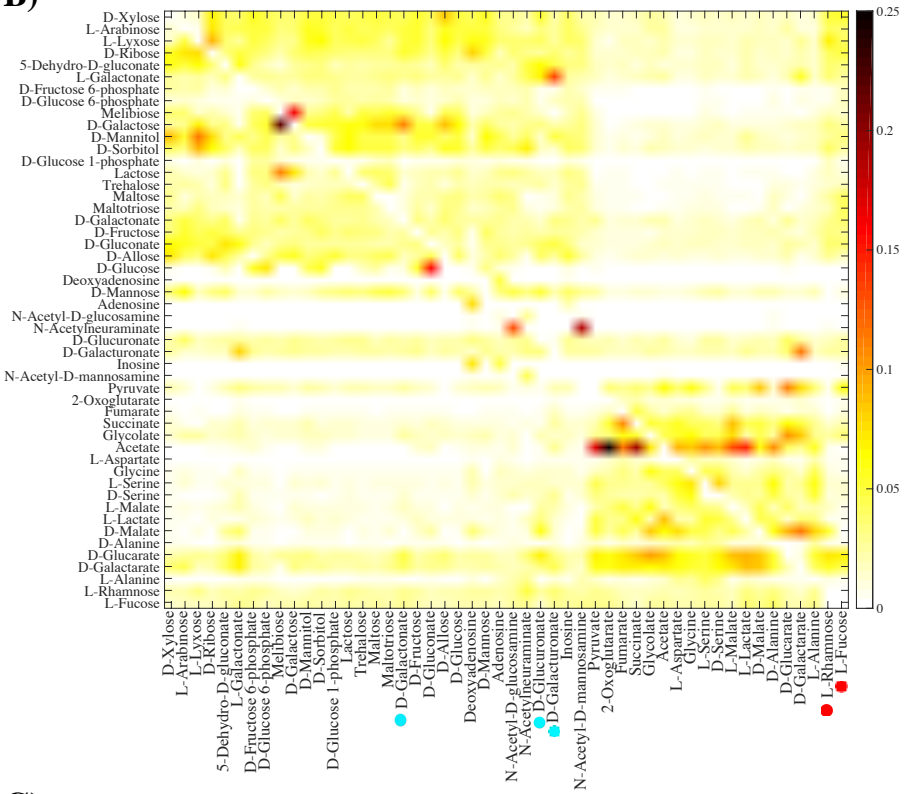


**Figure S18: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes. A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis. Recombinants are generated between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, with the exception of the gluconeogenic carbon sources D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), and the glycolytic carbon sources L-rhamnose, and L-fucose (shown by red circles). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks during recombination.

**A)**


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Xylose
L-Arabinose
L-Lyxose
D-Ribose
5-Dehydro-D-gluconate
L-Galactonate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Melibiose
D-Galactose
D-Mannitol
D-Sorbitol
D-Glucose 1-phosphate
Lactose
Trehalose
Maltose
Maltotriose
D-Galactonate
D-Fructose
D-Gluconate
D-Allose
D-Glucose
Deoxyadenosine
D-Mannose
Adenosine
N-Acetyl-D-glucosamine
N-Acetylneuraminate
D-Glucuronate
D-Galacturonate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Fumarate
Succinate
Glycolate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Malate
L-Lactate
D-Malate
D-Alanine
D-Glucarate
D-Galactarate
L-Alanine
L-Rhamnose
L-Fucose
0.25
0.2
0.15
0.1
0.05
0

**C)**

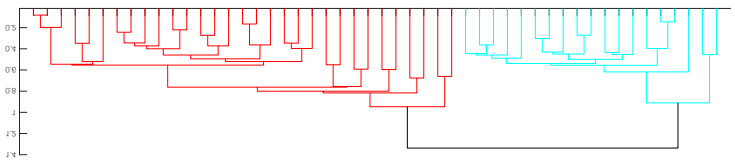
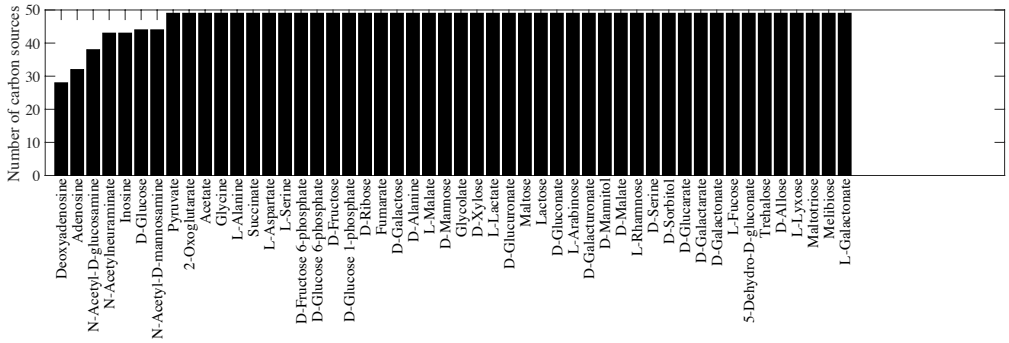
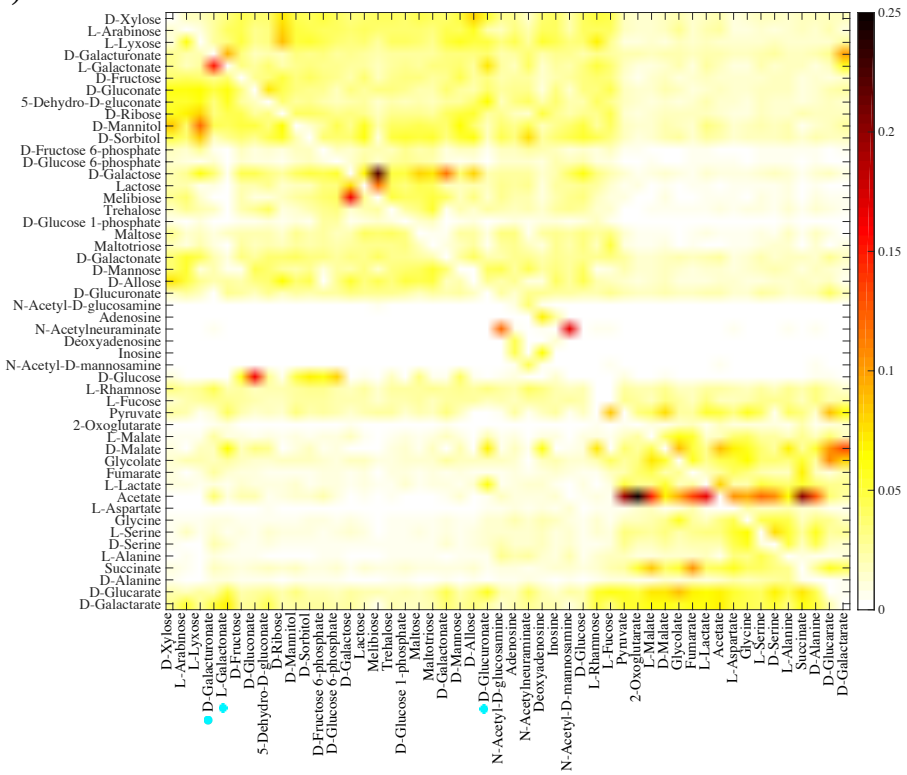


**Figure S19: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes ($n$ = 20). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 20 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
D-Glucose
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Xylose
L-Arabinose
L-Lyxose
D-Galacturonate
L-Galactonate
D-Fructose
D-Gluconate
5-Dehydro-D-gluconate
D-Ribose
D-Mannitol
D-Sorbitol
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Galactose
Lactose
Melibiose
Trehalose
D-Glucose 1-phosphate
Maltose
Maltotriose
D-Galactonate
D-Mannose
D-Allose
D-Glucuronate
N-Acetyl-D-glucosamine
Adenosine
N-Acetylneuraminate
Deoxyadenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
L-Rhamnose
L-Fucose
Pyruvate
2-Oxoglutarate
L-Malate
D-Malate
Glycolate
Fumarate
L-Lactate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Alanine
Succinate
D-Alanine
D-Glucarate
D-Galactarate
0.25
0.2
0.15
0.1
0.05
0

**C)**

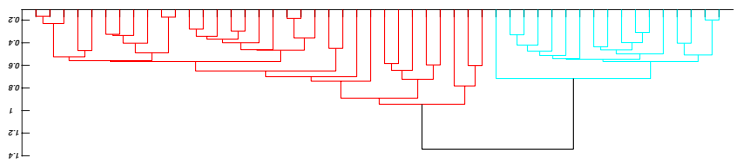
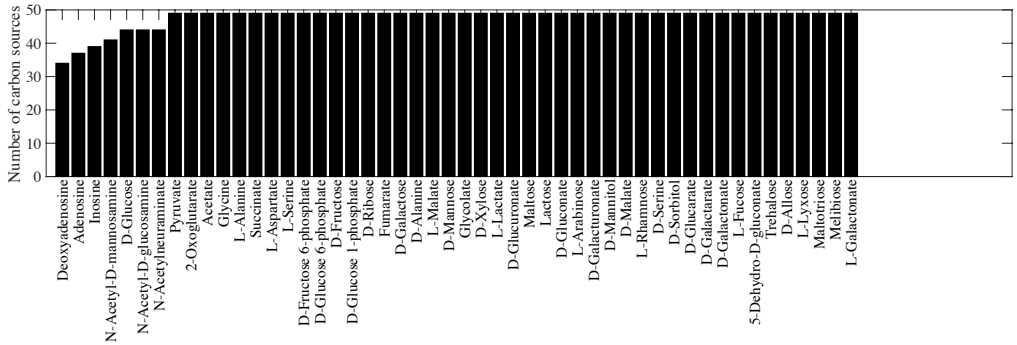
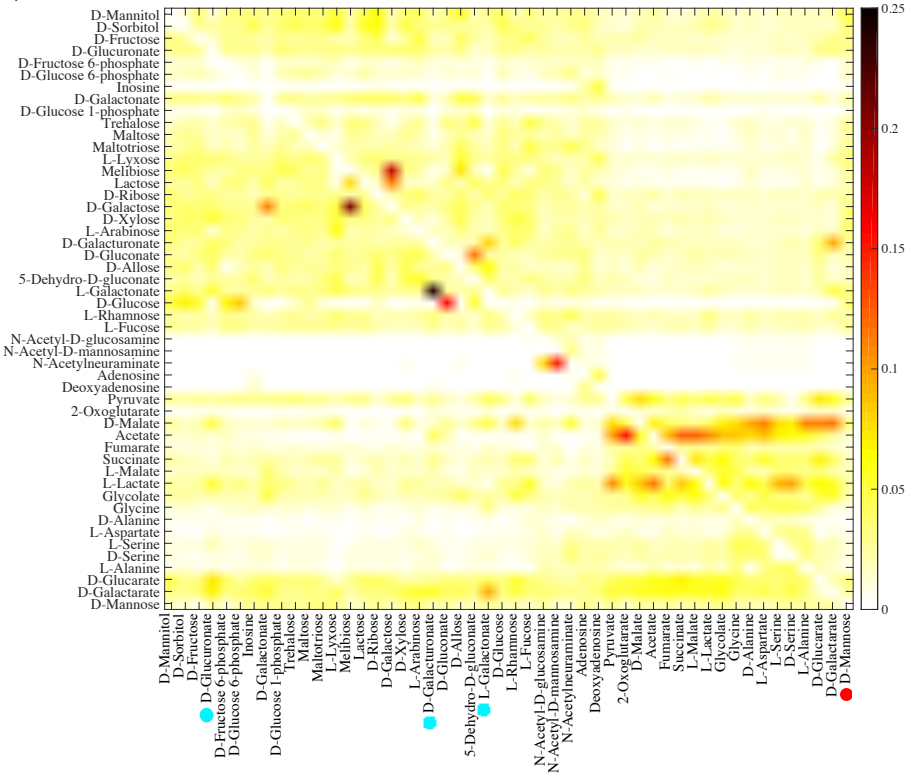

0.2
0.4
0.6
0.8
1
1.2
1.4

**Figure S20: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes ($n$ = 30). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 100 reactions. Moreover, $n$ = 30 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Mannitol
D-Sorbitol
D-Fructose
D-Glucuronate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Inosine
D-Galactonate
D-Glucose 1-phosphate
Trehalose
Maltose
Maltotriose
L-Lyxose
Melibiose
Lactose
D-Ribose
D-Galactose
D-Xylose
L-Arabinose
D-Galacturonate
D-Gluconate
D-Allose
5-Dehydro-D-gluconate
L-Galactonate
D-Glucose
L-Rhamnose
L-Fucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
N-Acetylneuraminate
Adenosine
Deoxyadenosine
Pyruvate
2-Oxoglutarate
D-Malate
Acetate
Fumarate
Succinate
L-Malate
L-Lactate
Glycolate
Glycine
D-Alanine
L-Aspartate
L-Serine
D-Serine
L-Alanine
D-Glucarate
D-Galactarate
D-Mannose
0.25
0.2
0.15
0.1
0.05
0

**C)**

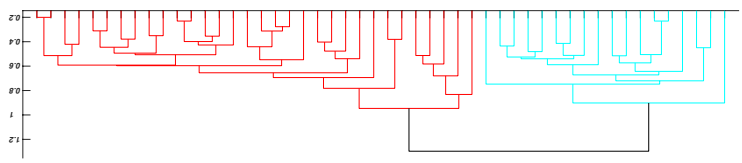
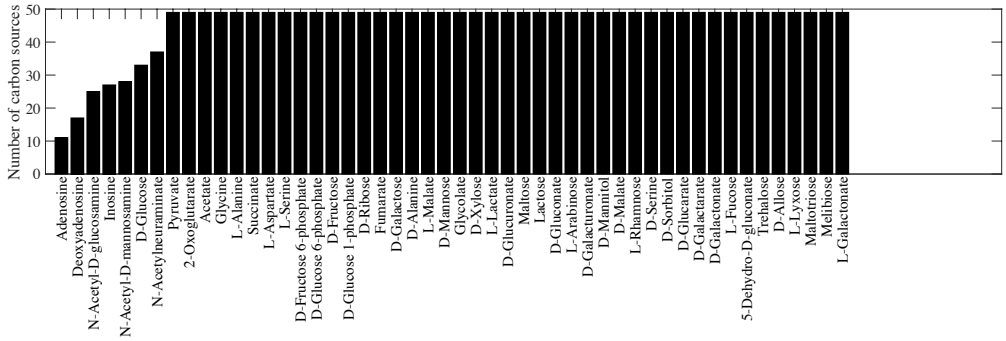
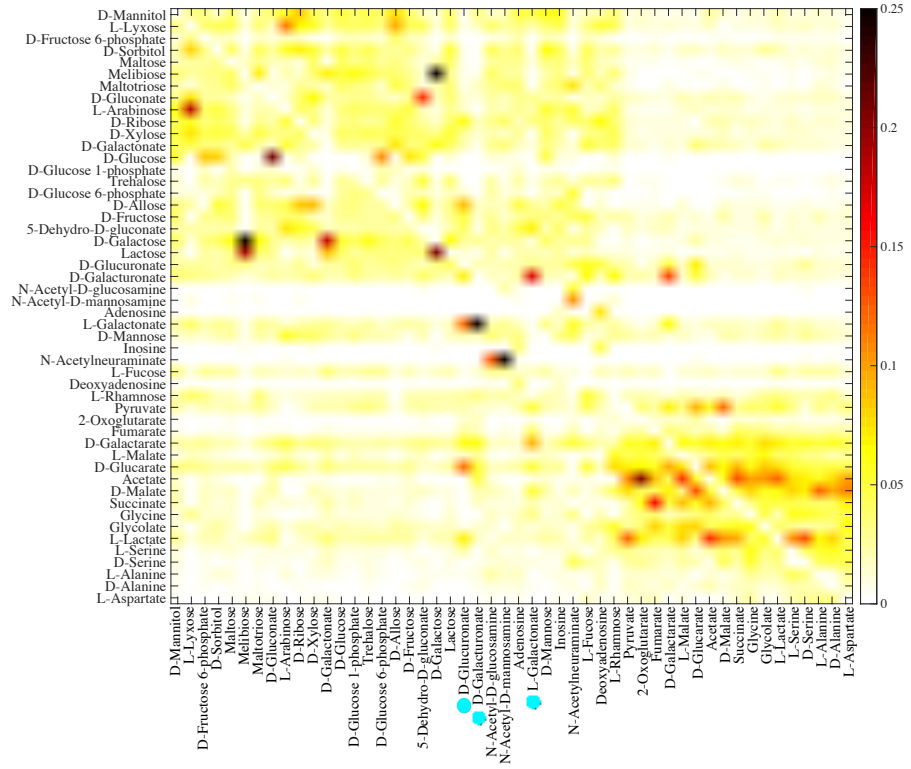

1.2
1
0.8
0.6
0.4
0.2

**Figure S21: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes ($D$ = 1,000). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and D-mannose (shown by red circles), which is a glycolytic carbon source). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E. coli* metabolic network, and they differ in $D$ = 1,000 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Mannitol
L-Lyxose
D-Fructose 6-phosphate
D-Sorbitol
Maltose
Melibiose
Maltotriose
D-Gluconate
L-Arabinose
D-Ribose
D-Xylose
D-Galactonate
D-Glucose
D-Glucose 1-phosphate
Trehalose
D-Glucose 6-phosphate
D-Allose
D-Fructose
5-Dehydro-D-gluconate
D-Galactose
Lactose
D-Glucuronate
D-Galacturonate
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
L-Galactonate
D-Mannose
Inosine
N-Acetylneuraminate
L-Fucose
Deoxyadenosine
L-Rhamnose
Pyruvate
2-Oxoglutarate
Fumarate
D-Galactarate
L-Malate
D-Glucarate
Acetate
D-Malate
Succinate
Glycine
Glycolate
L-Lactate
L-Serine
D-Serine
L-Alanine
D-Alanine
L-Aspartate
0.25
0.2
0.15
0.1
0.05
0

**C)**

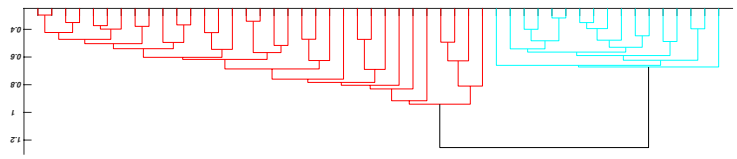
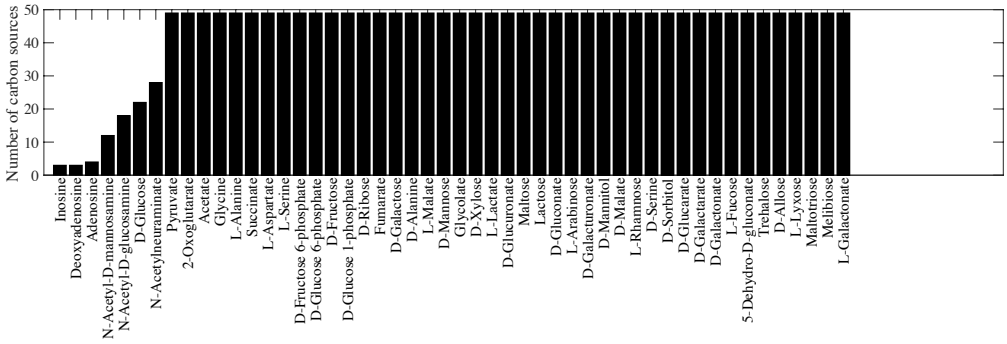
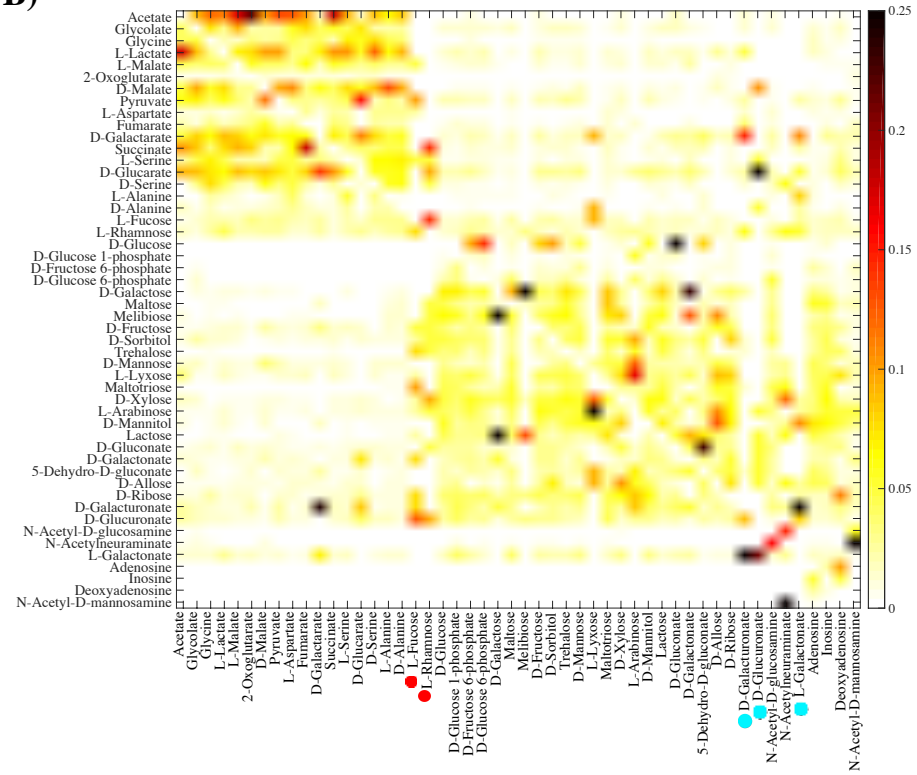

1.2
1
0.8
0.6
0.4

**Figure S22: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes ($||G||$ = 1,800). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources). In these analyses, parental metabolic networks contain $||G||$ = 1,800 reactions, and they differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
Inosine
Deoxyadenosine
Adenosine
N-Acetyl-D-mannosamine
N-Acetyl-D-glucosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


Acetate
Glycolate
Glycine
L-Lactate
L-Malate
2-Oxoglutarate
D-Malate
Pyruvate
L-Aspartate
Fumarate
D-Galactarate
Succinate
L-Serine
D-Glucarate
D-Serine
L-Alanine
D-Alanine
L-Fucose
L-Rhamnose
D-Glucose
D-Glucose 1-phosphate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Galactose
Maltose
Melibiose
D-Fructose
D-Sorbitol
Trehalose
D-Mannose
L-Lyxose
Maltotriose
D-Xylose
L-Arabinose
D-Mannitol
Lactose
D-Gluconate
D-Galactonate
5-Dehydro-D-gluconate
D-Allose
D-Ribose
D-Galacturonate
D-Glucuronate
N-Acetyl-D-glucosamine
N-Acetylneuraminate
L-Galactonate
Adenosine
Inosine
Deoxyadenosine
N-Acetyl-D-mannosamine
0.25
0.2
0.15
0.1
0.05
0

**C)**

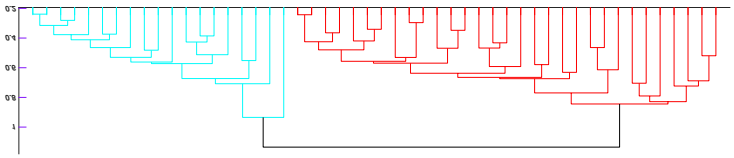
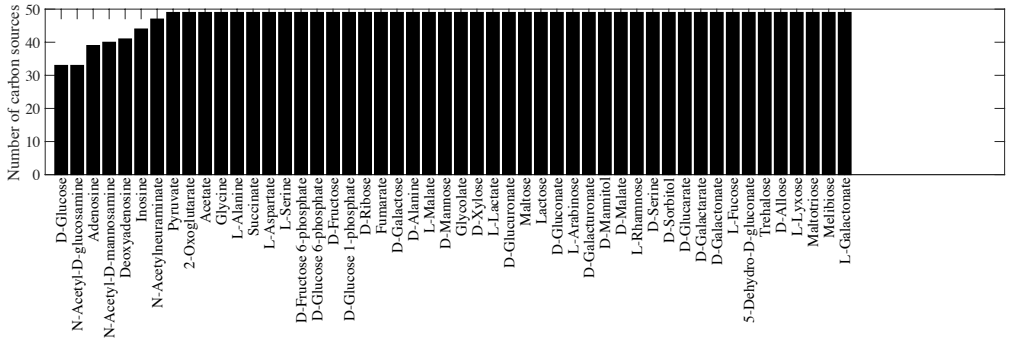
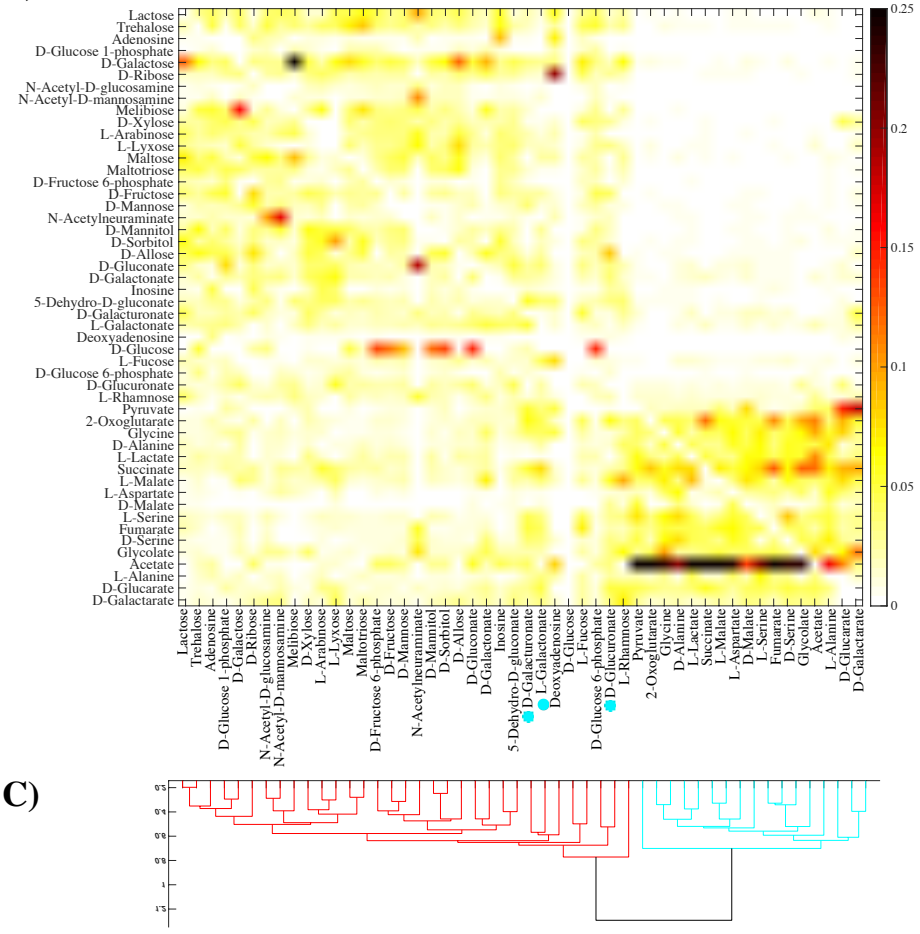


**Figure S23: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes ($||G||$ = 1,600). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). . In these analyses, parental metabolic networks contain $||G||$ = 1,600 reactions, and they differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
D-Glucose
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

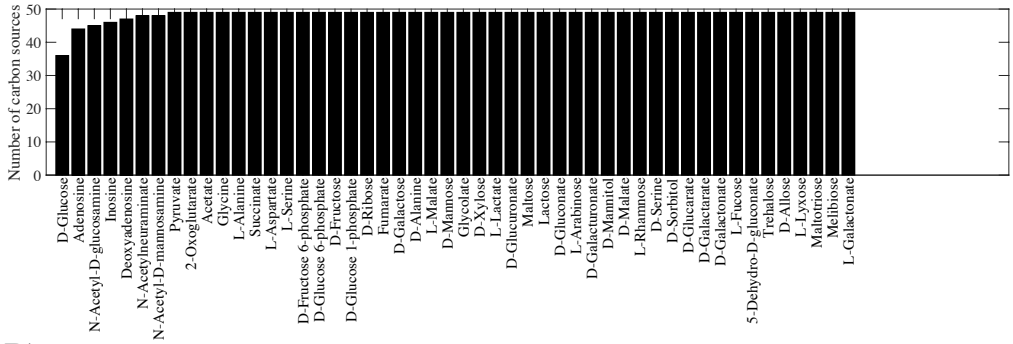
**B)**


Lactose
Trehalose
Adenosine
D-Glucose 1-phosphate
D-Galactose
D-Ribose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Melibiose
D-Xylose
L-Arabinose
L-Lyxose
Maltose
Maltotriose
D-Fructose 6-phosphate
D-Fructose
D-Mannose
N-Acetylneuraminate
D-Mannitol
D-Sorbitol
D-Allose
D-Gluconate
D-Galacturonate
Inosine
5-Dehydro-D-gluconate
D-Galactonate
L-Galactonate
Deoxyadenosine
D-Glucose
L-Fucose
D-Glucose 6-phosphate
D-Glucuronate
L-Rhamnose
Pyruvate
2-Oxoglutarate
Glycine
D-Alanine
L-Lactate
Succinate
L-Malate
L-Aspartate
D-Malate
L-Serine
Fumarate
D-Serine
Glycolate
Acetate
L-Alanine
D-Glucarate
D-Galactarate
0.25
0.2
0.15
0.1
0.05
0

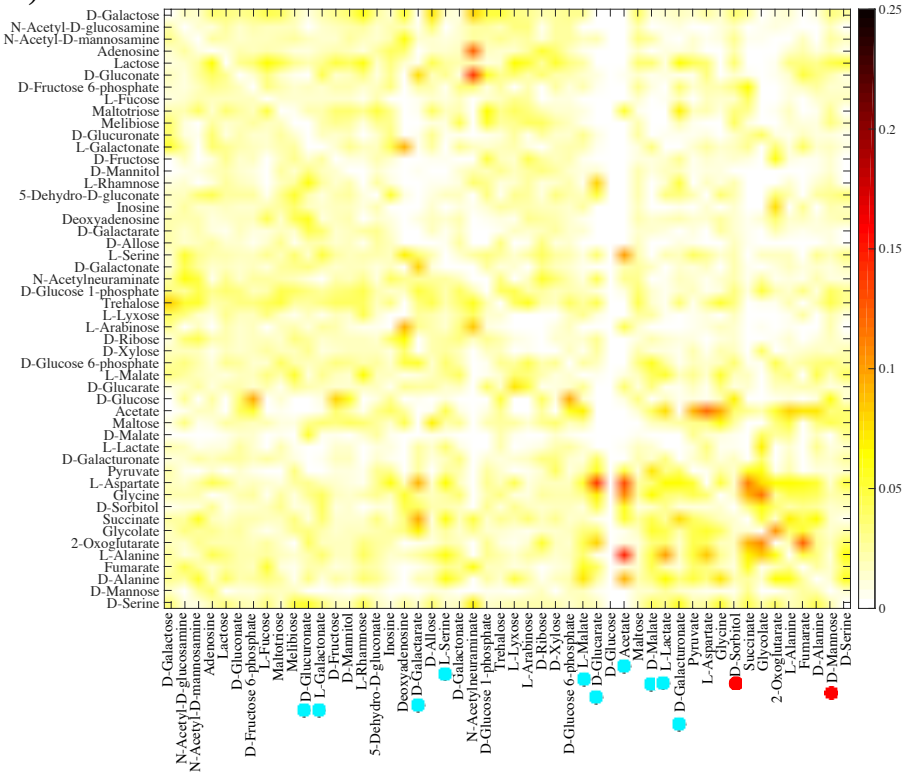
**C)**

**Figure S24: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (Parents with heterogeneous phenotypes, donors viable only on glucose). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between donor parents viable exclusively on glucose and the recipient parents that are exclusively viable on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 1,800 reactions, and differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

**A)**


Number of carbon sources
0
10
20
30
40
50
D-Glucose
Adenosine
N-Acetyl-D-glucosamine
Inosine
Deoxyadenosine
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

**B)**


D-Galactose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
Lactose
D-Gluconate
D-Fructose 6-phosphate
L-Fucose
Maltotriose
Melibiose
D-Glucuronate
L-Galactonate
D-Fructose
D-Mannitol
L-Rhamnose
5-Dehydro-D-gluconate
Inosine
Deoxyadenosine
D-Galactarate
D-Allose
L-Serine
D-Galactonate
N-Acetylneuraminate
D-Glucose 1-phosphate
Trehalose
L-Lyxose
L-Arabinose
D-Ribose
D-Xylose
D-Glucose 6-phosphate
L-Malate
D-Glucarate
D-Glucose
Acetate
Maltose
D-Malate
L-Lactate
D-Galacturonate
Pyruvate
L-Aspartate
Glycine
D-Sorbitol
Succinate
Glycolate
2-Oxoglutarate
L-Alanine
Fumarate
D-Alanine
D-Mannose
D-Serine
0.25
0.2
0.15
0.1
0.05
0

**C)**

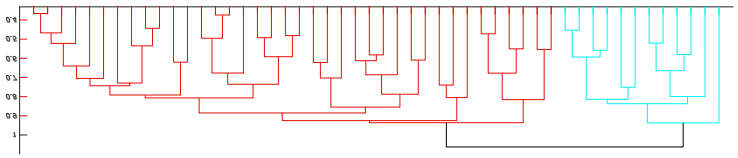


**Figure S25: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (Parents with heterogeneous phenotypes, recipients viable only on glucose). A)** The horizontal axis shows the carbon use phenotype $C_i$ of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on $C_i$. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between recipient parents viable exclusively on glucose and donor parents that are exclusively viable on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (with 12 exceptions; shown by 10 cyan circles, and 2 red circles.). In these analyses, parental metabolic networks contain $||G||$ = 1,800 reactions, and differ in $D$ = 100 reactions. Moreover, $n$ = 10 reactions are swapped between parental metabolic networks in a recombination event.

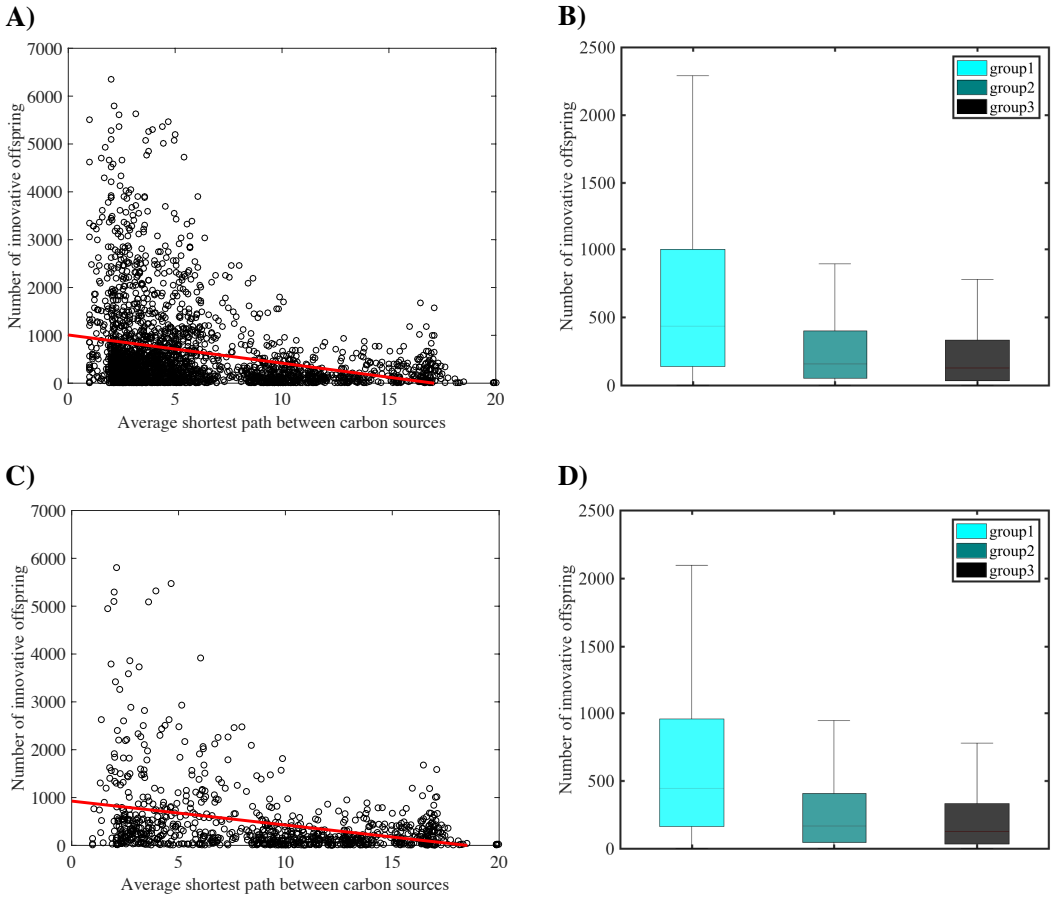



**Figure S26: Distance between carbon sources in substrate graphs and relative constraint in the emergence of innovative offspring.** In all 4 panels, the vertical axis shows the number of innovative recombinants (per 1 million recombinant offspring) gaining viability on some new carbon source $C_j$ resulting from recombination between parental metabolic networks viable on carbon source $C_i$. In panels A and C, the horizontal axes show the mean shortest path between carbon source $C_i$ and $C_j$ in the substrate graph (supplementary text S7) of the metabolic networks viable on carbon source $C_i$. In panel **A)** each circle corresponds to a given pair of carbon sources ($C_i$, $C_j$), and data on both axes are significantly correlated (Pearson $r$=*-0.2722*, and $P<10^{-41}$). In panel **B)** the carbon source pairs ($C_i$, $C_j$) are divided into three groups based on their mean shortest path ($||SP\ (i,j)||$) between carbon source $C_i$ and $C_j$ in the substrate graph of metabolic networks viable on carbon source $C_i$: group 1 $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), group 2 $\{i,j|6 < ||SP\ (i,j)|| \leq 1\}$), and group 3 $\{i,j|\ ||SP\ (i,j)|| > 12\}$. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.

In panel A, a non-uniform distribution of mean shortest paths ($||SP\ (i,j)||$) between carbon sources is evident on the horizontal axis. To exclude the possibility that the correlation in panel A is significant simply because of a higher number of data points for lower shortest path distances, we repeated the analyses shown in panels A and B by resampling from the

2500 pairs of carbon sources an equal number of pairs in each distance category, i.e., 284 pairs ($C_i$, $C_j$) with $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), 284 pairs ($C_i$, $C_j$) with $\{i,j|6 < ||SP\ (i,j)|| \leq 12\}$), and 284 pairs ($C_i$, $C_j$) with $\{i,j|\ ||SP\ (i,j)|| > 12\}$, to create the subsampled data in panels C and D. In panel **C)** each circle corresponds to a given pair of carbon sources ($C_i$, $C_j$), and data on both axes are significantly correlated (Pearson *r=-0.3411*, and $P<10^{-24}$). In panel **D)**, analogous to panel B, carbon source pairs ($C_i$, $C_j$) are divided into three equally-sized groups based on their mean shortest path ($||SP\ (i,j)||$) between carbon source $C_i$ and $C_j$ in the substrate graph of metabolic networks viable on carbon source $C_i$: group 1 $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), group 2 $\{i,j|6 < ||SP\ (i,j)|| \leq 1\}$), and group 3 $\{i,j|\ ||SP\ (i,j)|| > 12\}$. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E. coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

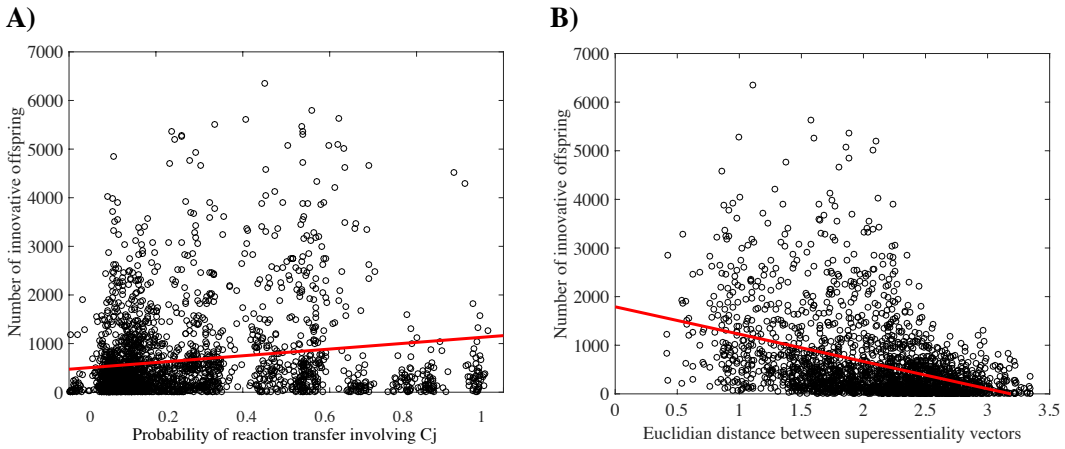



**Figure S27:** In both panels, each circle corresponds to a given pair of carbon sources ($C_i$, $C_j$) and the vertical axis shows the number of innovative recombinants (per 1 million recombinant offspring) gaining viability on some new carbon source $C_j$ resulting from recombination between parental metabolic networks viable on carbon source $C_i$. The horizontal axes show **A)** the fraction of parental metabolic network pairs viable on carbon source $C_i$, in which a reaction that can enable viability on carbon source $C_j$ can be transferred from the donor to the recipient metabolic network, and **B)** the Euclidian distance between superessentiality vectors of the corresponding pair of carbon sources, which we use as another proxy for the biochemical distance between carbon sources. In both panels the data plotted against one another are significantly correlated: **A)** Pearson *r=0.163*, and $P<10^{-15}$, and **B)** Pearson *r=-0.3935*, and $P<10^{-83}$. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E. coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic

networks during recombination.

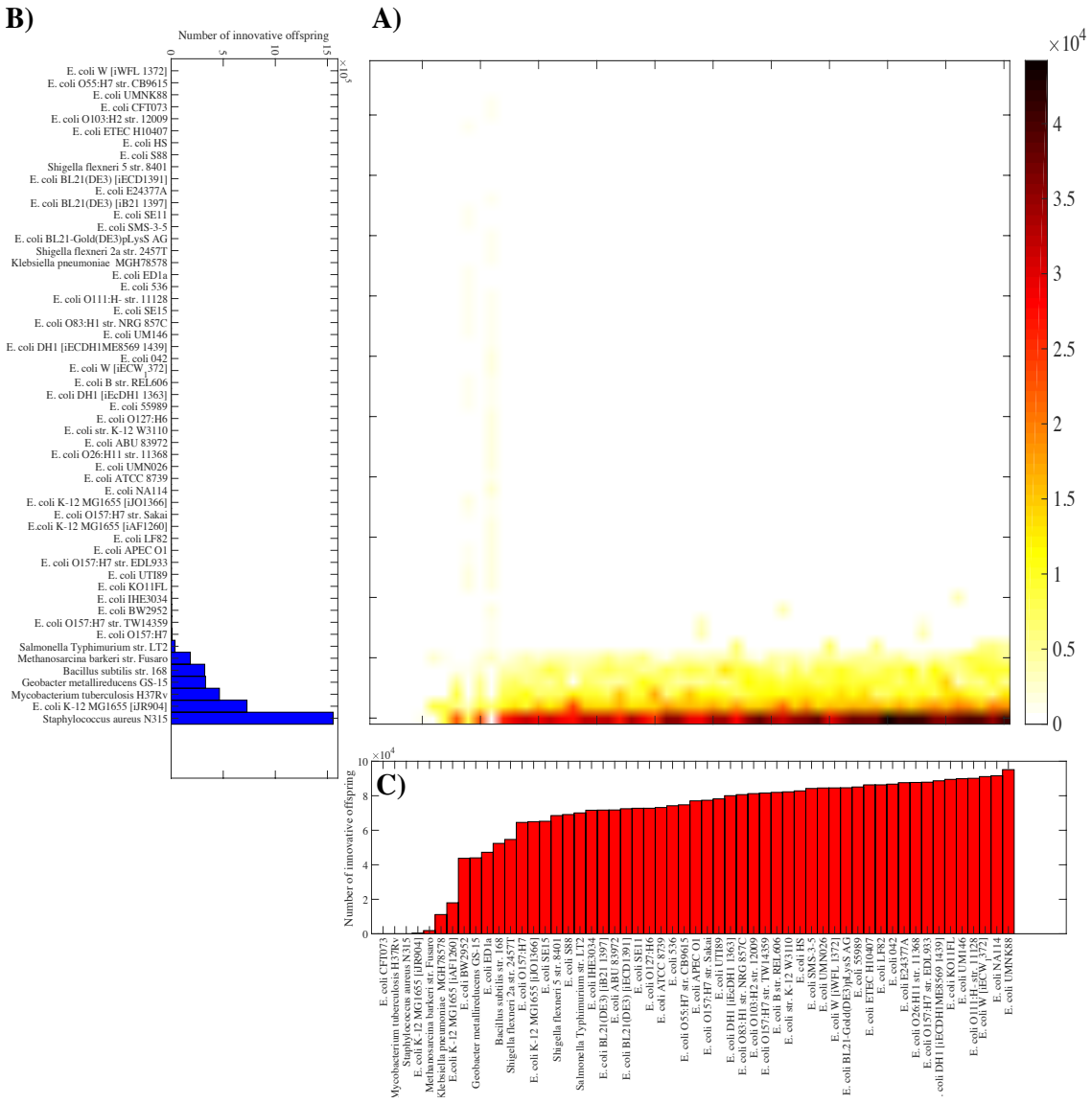



**Figure S28: Emergence of innovative offspring is contingent on and constrained by parental genotypes. A)** Number of innovative offspring resulting from linkage-based recombination between bacterial DNA donors specified on the vertical axis of panel B, and the corresponding recipient genotypes specified on the horizontal axis of panel C (coded according to the color legend). **B)** Total number of innovative recombinant offspring involving the donor genotype specified on the vertical axis. **C)** Total number of innovative recombinant offspring involving the recipient genotype specified on the horizontal axis.

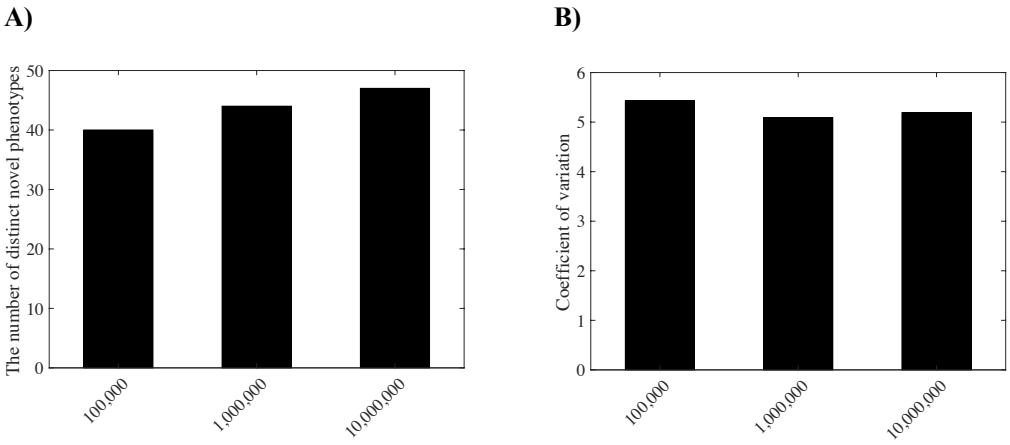



**Figure S29: Sample size and its effect on absolute and relative constraints.** For this analysis, we used 1,000 parental metabolic networks that are viable exclusively on glucose, and in three different simulations we generated *i)* 100, *ii)* 1,000 and *iii)* 10,000 offspring from each parent, which amounts to *i)* 100,000 *ii)* 1,000,000 and *iii)* 10,000,000 total offspring, as indicated on the horizontal axes. The vertical axes show **A)** the number of distinct novel phenotypes (among a possible total of 49 phenotypes) that emerged in the offspring, and **B)** the coefficient of variation in the number of innovative offspring for different novel carbon usage phenotypes. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E. coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

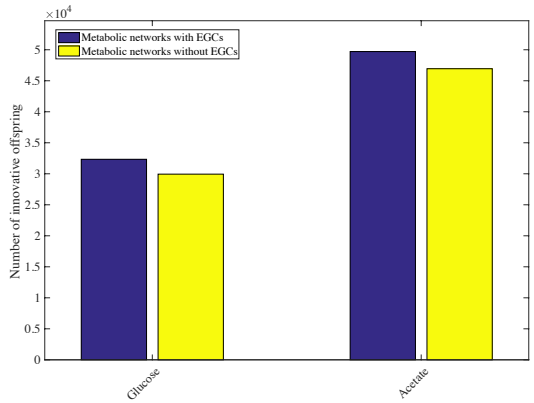
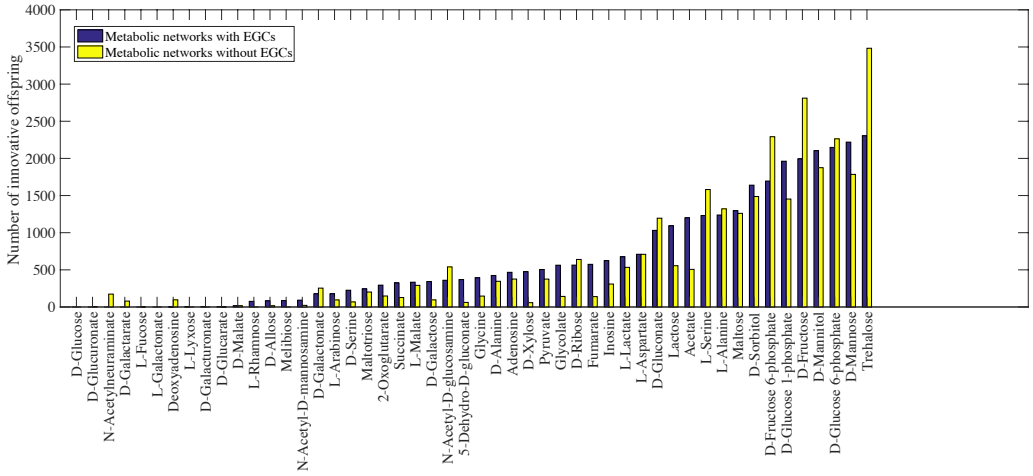



**Figure S30: Erroneous energy generating cycles (EGCs) and the emergence of innovative offspring.** The number of innovative offspring (per 1 million recombinants) emerging from recombination between parental metabolic networks that contain EGCs (blue) or that do not contain EGCs (yellow), and that are viable exclusively on glucose (left) and acetate (right). In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E. coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

**A)**




**B)**

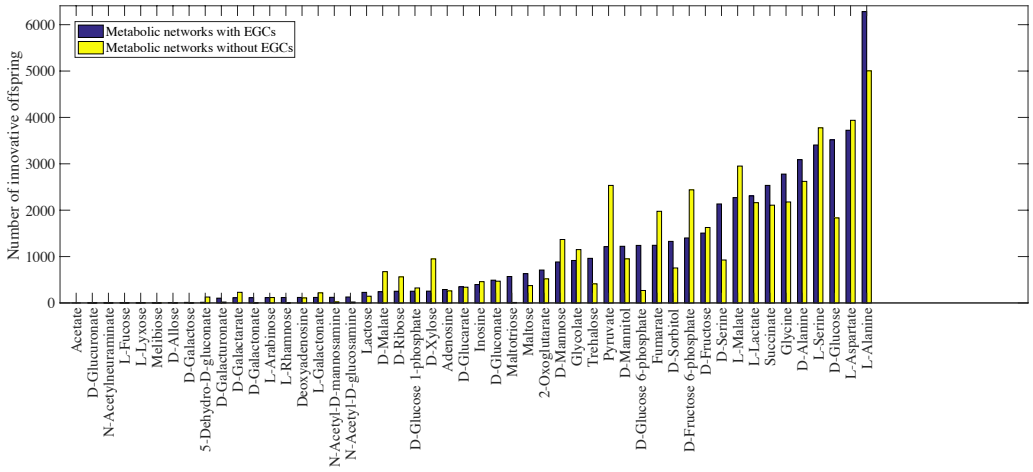



**Figure S31: Erroneous energy generating cycles (EGCs) and relative constraints.** Horizontal axes show the number of innovative offspring (per 1 million recombinants) emerging from recombination between parental metabolic networks viable exclusively on **A)** glucose and **B)** acetate, where parental metabolisms contain EGCs (blue) or do not contain EGCs (yellow). The ranking of the height of the blue bars and yellow bars in both panels is significantly correlated (panel A: Spearman's $\rho = 0.8913$, and $P < 10^{-18}$; panel B: Spearman's $\rho = 0.9197$, and $P < 10^{-21}$). In these analyses, parental metabolic networks contain $||G||=2079$ reactions, the same as the *E. coli* metabolic network, and they differ in $D=100$ reactions. Moreover, $n=10$ reactions are swapped between parental metabolic networks during recombination.

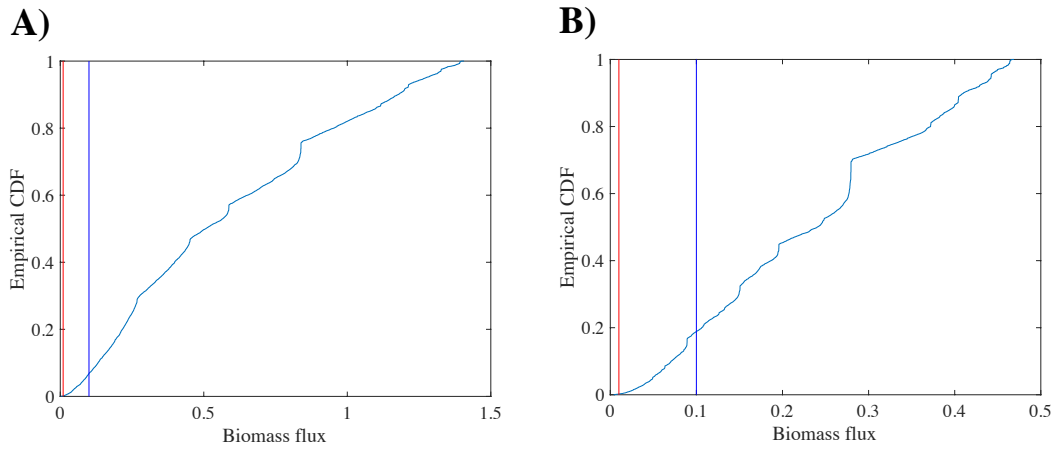



**Figure S32: Biomass growth flux of most viable metabolic networks is much greater than our cut-off value for viability.** The vertical axes show the empirical cumulative distribution function of the biomass flux among 10,000 MCMC-sampled metabolic networks viable exclusively on **A)** glucose, and **B)** acetate. The vertical red and blue lines show the cut-off value of 0.01 and 0.1 $1/h$. We used 0.001 $1/h$ as the cut-off value for viability.

# Chapter 5:

# Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms

Aditya Barve, Sayed-Rzgar Hosseini, Olivier C Martin and Andreas Wagner

The content of this chapter has been published as:

Barve, A., S.-R. Hosseini, O.C. Martin, and A. Wagner. 2014. Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. BMC Syst. Biol. 8: 48.
https://doi.org/10.1186/1752-0509-8-48

Note: Although I am not considered as the leading author of this paper, I have made substantial contribution on this manuscript. I have performed all the analyses related to central carbon metabolism that accounts for more than 75% of the content of this paper. I developed an algorithm that efficiently determines the connected components of large metabolic genotype networks. Moreover, I suggested the inference of connectivity of very large genotype networks based on parent-child relationships of genotype networks of consecutive size (see figures 2 and 3 in the manuscript), which made the entire analysis possible. However, I did not have a direct role in writing the manuscript.

## 5.1. Abstract

A metabolism can evolve through changes in its biochemical reactions that are caused by processes such as horizontal gene transfer and gene deletion. While such changes need to preserve an organism's viability in its environment, they can modify other important properties, such as a metabolism's maximal biomass synthesis rate and its robustness to genetic and environmental change. Whether such properties can be modulated in evolution depends on whether all or most viable metabolisms – those that can synthesize all essential biomass precursors – are connected in a space of all possible metabolisms. Connectedness means that any two viable metabolisms can be converted into one another through a sequence of single reaction changes that leave viability intact. If the set of viable metabolisms is disconnected and highly fragmented, then historical contingency becomes important and restricts the alteration of metabolic properties, as well as the number of novel metabolic phenotypes accessible in evolution. We here computationally explore two vast spaces of possible metabolisms to ask whether viable metabolisms are connected. We find that for all but the simplest metabolisms, most viable metabolisms can be transformed into one another by single viability-preserving reaction changes. Where this is not the case, alternative essential metabolic pathways consisting of multiple reactions are responsible, but such pathways are not common. Metabolism is thus highly evolvable, in the sense that its properties could be fine-tuned by successively altering individual reactions. Historical contingency does not strongly restrict the origin of novel metabolic phenotypes.

## 5.2. Introduction

For biological systems on different levels of organization, the same broadly defined phenotype can usually be formed by more than one genotype. Examples include RNA, where many genotypes (sequences) share the same secondary structure phenotype [1, 2, 3, 4]; proteins, where multiple amino acid sequences form the same fold [5, 6]; regulatory circuits, where many genetically encoded circuit topologies can form the same expression pattern [7, 8, 9]; and metabolism, where multiple metabolic genotypes, encoding different combinations of chemical reactions, can confer viability on the same spectrum of nutrients [10, 11, 12, 13]. The number of genotypes with the same phenotype is usually astronomical. For example, it can exceed $10^{20}$ for moderately long RNA molecules of 40 nucleotides with the same secondary structure [14]; it has been estimated at $10^{57}$ for proteins that adopt a fold characteristic of the bacteriophage λ transcriptional repressor [15], and at more than $10^{40}$ for model regulatory circuits of 10 genes that form a given gene expression pattern [7].

The many different genotypes that share one aspect of their phenotype may differ in other aspects, such as the thermodynamic stability of a given RNA or protein fold, the resilience of a gene expression pattern to stochastic noise, or the robustness of a metabolism to deletion of genes that encode metabolic enzymes [1, 7, 16, 17]. Because such properties can be important for the biological function of any one system, the question whether they can be "fine-tuned" in evolution is important [7, 18, 19, 20]. Such fine-tuning may depend on whether one can start from any one genotype with a given phenotypic property and reach most other such genotypes through sequences of small genetic change.

Whether such fine-tuning is possible can be studied in the framework of a space of possible genotypes, where two genotypes are adjacent if they differ by the smallest possible genetic change, such as a single amino acid change in two proteins. In this framework, the question becomes whether a set of genotypes with the same phenotype forms a single connected genotype network (also known as a neutral network [1]), or whether this network fragments into multiple isolated subnetworks or *disconnected components* [21].

Whenever such fragmentation occurs, the constraint it imposes on genotypic change does not only affect the ability to modulate a phenotype. It also gives an important role to historical accidents in the evolutionary process: The genotype with a given phenotype that evolution happened to have "discovered" first can determine the number and identity of other genotypes reachable through gradual genetic change. And by restricting the number of accessible genotypes, fragmentation can also restrict the spectrum of novel phenotypes accessible as new adaptations. The reason is that this spectrum depends strongly on a genotype's location in genotype space [22]. The further evolution can "walk away" from a given genotype, the more the spectrum of accessible phenotypes changes [1, 11, 23, 24, 25, 26]. In sum, fragmentation of a genotype network can cause historical contingency and restrict a system's potential for future evolutionary change.

Existing work, based on computational models of phenotype formation, shows that fragmentation is system-dependent. For example, in RNA secondary structure phenotypes, genotype networks are typically highly fragmented [18, 27], whereas for regulatory circuits, such fragmentation depends on the kind of circuit studied, its size, and how one defines its gene expression phenotypes [7, 8, 28]. Because the question has thus far not been answered in metabolic systems, we here analyze the connectedness of a space of metabolisms.

A metabolism is a complex network of chemical reactions, catalyzed by enzymes and encoded by genes, whose most fundamental task is to synthesize multiple small molecule precursors for biomass, such as amino acids, nucleotides, and lipids [29, 30]. An organism's metabolic genotype is the part of a genome that encodes metabolic genes. It is thus fundamentally a string of DNA, but can be represented more compactly as a binary vector of length $N$, where $N$ is the number of metabolic reactions in a known "universe" of metabolic reactions (Additional file 1[10, 11], see Methods). This universe comprises all enzyme-catalyzed reactions known to take place in some organism. The $i$-th entry of this vector corresponds to the $i$-th reaction in a list of such reactions, and for any one organism, the value of this entry is one if the organism can catalyze the $i$-th reaction, and zero otherwise. On evolutionary time scales, the reaction complement of a metabolism can change through processes such

as horizontal transfer of enzyme-coding genes, gene deletions, as well as gene duplications followed by sequence divergence.

The known "universe" of metabolism currently comprises more than $N = 5000$ reactions [31, 32]. This means that there are more than $2^{5000}$ different metabolic genotypes, which constitute a vast space of possible metabolisms. For any one metabolism in this space and any one chemical environment, one can compute the spectrum of biomass precursors that it can synthesize using the constraint-based computational method of flux-balance analysis (FBA). We call any one metabolism *viable* in a specific chemical environment, if it can synthesize every single one in a spectrum of essential biomass precursors from nutrients in this environment [10, 11, 13, 33] (see Methods). We will here consider minimal chemical environments that contain only one carbon source, such as glucose, as the sole carbon source.

Because connectedness of a metabolic genotype network may depend on the number $n$ of reactions in a metabolism, we distinguish in our analysis metabolisms of different sizes. If $\Omega(n)$ is the set of all metabolisms with $n$ biochemical reactions ($n \leq N$) and if $V(n)$ is the subset of all viable metabolisms, we are interested in whether $V(n)$ is connected. Because the metabolisms of free-living heterotrophic metabolisms may have thousands of reactions, we need to study $V(n)$ for metabolisms this large. This is not an easy task, because the set of viable metabolisms is so enormous that exhaustive enumeration is impossible [10, 34]. Therefore, to sharpen our intuition and to illustrate key concepts, we first analyze a smaller metabolic genotype space whose viable metabolisms can be enumerated exhaustively. This is the space of metabolisms that can be formed by subsets of $N = 51$ reactions in central carbon metabolism [35] (see Methods). Even though central carbon metabolism is highly conserved, its reaction complement varies in nature, for example through variants of glycolysis [36, 37, 38, 39, 40] and the tricarboxylic acid cycle, where some organisms have an incomplete cycle [41]. We go beyond such naturally occurring variation and analyze metabolisms comprised of all possible subsets of all 51 reactions. Even though this number of metabolisms is astronomical ($2^{51} \approx 10^{15}$), we were able to determine viability for all of them, and thus analyze the connectivity of $V(n)$ for all $n \leq N$ ($N = 51$). After that, we turn to larger, genome-scale metabolisms, where we study the connectivity of $V(n)$ through a sampling approach.

As many of the metabolisms used in our analysis may not be realized in extant organisms, we also refer to them as potential metabolisms.

Our observations show that for all but the simplest metabolisms, those that contain close to the minimal number of reactions necessary for viability, most viable potential metabolisms $V(n)$ lie on a single connected genotype network. Where fragmentation into different components occurs, its biochemical cause are alternative biochemical pathways that occur in different components, that are essential for the synthesis of specific biomass precursors, that comprise more than one reaction, and that cannot be transformed into one another by changes in single reactions without destroying viability. Because such pathways only occur in the smallest potential metabolisms, fragmentation and thus historical contingency do not strongly constrain the evolution of properties such as robustness, biomass synthesis rate, or the accessibility of novel metabolic phenotypes.

## 5.3. Results

### Study System 1: Central Carbon Metabolism

Our first analysis focuses on potential metabolic genotypes that can be formed with subsets of $N = 51$ reactions in the central carbon metabolism of *E. coli* [35] (see Methods). This metabolic core of *E. coli* includes reactions from glycolysis/gluconeogenesis, the tricarboxylic acid cycle, oxidative phosphorylation, pyruvate metabolism, the pentose phosphate shunt, as well as some reactions from glutamate metabolism (Additional file 2). It produces 13 precursor molecules (Additional file 2) that are required to synthesize all 63 small biomass molecules of *E. coli*, including nucleotides, amino acids, and lipids [30, 35, 42]. Examples of these precursors include oxaloacetate, a metabolite participating in the tricarboxylic acid cycle, which is used in the synthesis of amino acids such as asparagine, aspartate, lysine, and threonine [29, 42]. Another example is ribose-5-phosphate, which participates in the pentose phosphate pathway, and is necessary for the synthesis of nucleotides and amino acids, such as histidine, phenylalanine, and tryptophan [29, 42]. In our analysis, we consider a metabolism *viable* only if it can synthesize all 13 of these biomass precursors in a well-defined minimal environment containing a specific *sole* carbon source, such as glucose (see Methods).

**The fraction of viable genotypes is extremely small and decreases as metabolism size *n* decreases**

For each $n \leq N = 51$, we here explore the space $\Omega(n)$ of metabolisms (metabolic genotypes) with a given number of $n$ reactions. We represent each such metabolism as a binary vector of length $N = 51$, whose $i$-th entry is equal to one if the $i$-th reaction is present and zero otherwise. The largest metabolism ($n = N$) is the one where all reactions are present. The space of all possible metabolisms that contain a subset of these 51 reactions has $2^{51}$ ($\approx 10^{15}$) member genotypes, while for a given $n$, $\Omega(n)$ contains $\binom{51}{n}$ genotypes. We are especially interested in the subset $V(n)$ of $\Omega(n)$ that consists only of viable metabolisms. Because, $\Omega(n)$ can be very large, determining $V(n)$ is no small undertaking. For example, for metabolisms with $n = 30$, $\Omega(n)$ contains more than $1.14 \times 10^{14}$ genotypes, and the viability of each of them cannot be determined by brute force. However, one can use some peculiarities of metabolism to render this computation feasible (see Methods). For example, consider a metabolism (the "parent") with $n$ reactions and another metabolism (the "child") derived from it by deleting one reaction. If the parent is not viable then the child will not be viable either. By analyzing the viability of metabolisms with decreasing numbers of reactions $n$, and taking advantage of this relationship, we were able to reduce the computational cost of enumerating viable metabolisms by a factor $\approx 10^6$ to the evaluation of viability for only $1.55 \times 10^9$ metabolisms [43].

Figure 1A shows the number of viable metabolisms $V(n)$ (grey circles), together with the number of all metabolisms (black circles, $\Omega(n) = \binom{51}{n}$ as a function of the number $n$ of reactions. Note the logarithmic vertical axis. The number of viable metabolisms has a maximum at $n = 37$ with a total of $2.39 \times 10^8$ metabolisms, while the minimum size of a viable metabolism, i.e., the smallest $n$ such that $V(n) > 0$ is 23 (Additional file 3). This means that at least 23 reactions are required to synthesize all 13 biomass precursors on glucose. There are three such smallest metabolisms, one of which is shown in Additional file 4. Figure 1B expresses $V(n)$ as a fraction of the number of metabolisms $\Omega(n)$ (grey circles), and shows that this fraction decreases with decreasing $n$. This means that random sampling is much less likely to yield a viable metabolism for small than for large metabolisms. For the smallest $n$ with

viable metabolisms ($n = 23$), the three viable potential metabolisms correspond to a fraction $10^{-14}$ of all metabolisms of size 23. The largest viable metabolism contains all $n = 51$ reactions.

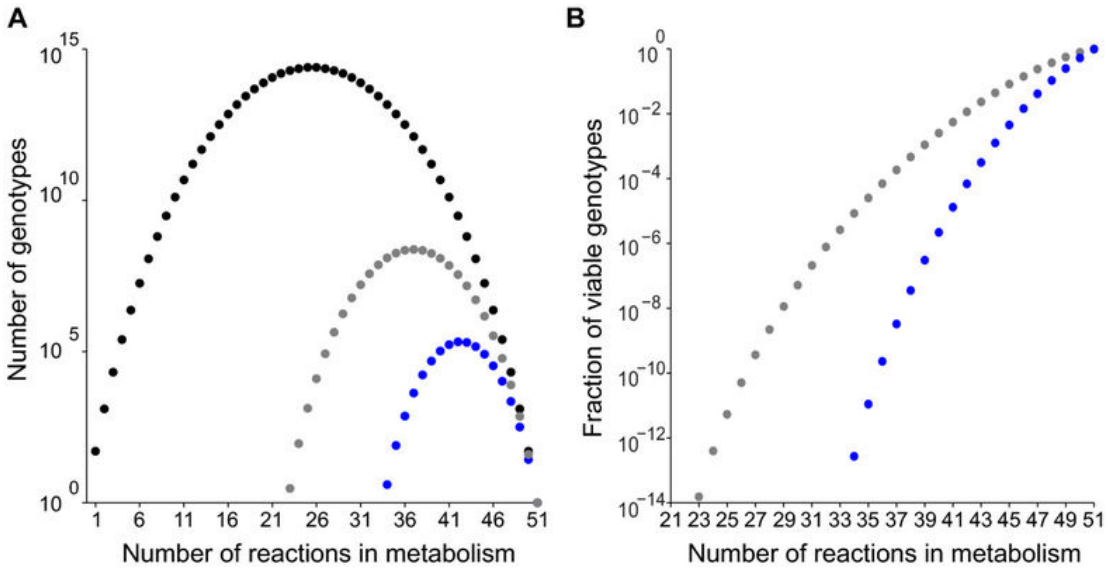



**Figure 1: The number of viable metabolisms** V ( n ) **decreases as the number of reactions** n **decreases. (A)** The vertical axis (note the logarithmic scale) shows the number of genotypes, and the horizontal axis shows the number $n$ of reactions in a potential metabolism. Black circles represent the number of genotypes in genotype space $\Omega(n)$ (regardless of viability), grey circles show the number of potential metabolisms viable on glucose, whereas the blue circles denote the number of potential metabolisms viable on all 10 carbon sources. **(B)** The vertical axis (note the logarithmic scale) shows the fraction $|V(n)| / |\Omega(n)|$. The grey circles show the fraction of genotypes viable on glucose relative to the number of possible metabolisms, whereas the blue circles denote the fraction of genotypes viable on 10 carbon sources relative to the number of possible metabolisms. Note that viable genotypes become extremely rare as the number of reactions in a metabolism decreases. Data for both figures is based on all viable metabolisms for each $n$ (Additional file 3 and Additional file 8).

## Useful principles to determine the connectedness of genotype networks

The viable genotypes at any one size $n$ can be represented as a genotype network, a graph whose nodes are genotypes, and where two genotypes are adjacent (connected by an edge), if they share all but one reaction. For example, the two hypothetical genotypes $G_1$ and $G_2$, where $G_1$ consists of reactions $\{R_1, R_2, R_3\}$, and $G_2$ consists of reactions $\{R_2, R_3, R_4\}$, are adjacent. This is because $G_1$ and $G_2$ share two out of the three reactions ($R_2$ and $R_3$). One can reach $G_2$ from $G_1$ by adding reaction $R_4$ and removing $R_1$, an event that we refer to as a reaction swap [10, 12, 33]. This

definition of neighboring genotypes allows us to keep the number of reactions in a genotype network constant. We note that each reaction swap can be decomposed into the addition of a reaction followed by the deletion of a reaction, both of which preserve viability provided that the reaction swap does. In other words, genotype networks that are connected if adjacency is defined under reaction swaps will remain connected if adjacency is defined via a sequence of alternating reaction additions and reaction deletions.

Our principal goal is to identify whether genotype networks at any one size $n$ are connected. This first requires us to establish the adjacency of $\binom{V(n)}{2}$ genotype pairs, followed by application of standard graph theory algorithms such as breadth-first search [21, 44] to compute whether genotypes decompose into two or more disconnected components, or whether they form a single connected network, i.e., whether a path through $V(n)$ exists connecting any two genotypes [21]. Because $V(n)$ exceeds $10^6$ genotypes at intermediate $n$ (Additional file 3), such conventional methods lead to large computational cost for all but the largest and smallest metabolisms ($n = 23–28$ and $n = 46–50$ reactions). For genotype networks comprising metabolisms of intermediate size ($n = 29–45$), we therefore took advantage of another relationship between "parent" and "child" metabolisms, namely that the connectivity of a genotype network at size $n$ can be understood based on its connectivity at size $n$-1. We explain this relationship next.

Starting from a genotype $G(n)$ with $n$ reactions, one can obtain a parent genotype $G(n + 1)$ with $(n + 1)$ reactions by adding to it any one reaction among the $N = 51$ reactions that are not already part of $G(n)$. Because addition of a reaction does not eliminate viability, $G(n + 1)$ will be viable, and thus be a member of $V(n + 1)$. For any one genotype $G(n)$, there exist $N$-$n$ reactions that are not part of this genotype. Therefore, one can obtain exactly $N$-$n$ genotypes of size $n + 1$ by adding a single reaction to a genotype $G(n)$. And because each pair of these genotypes of size $n + 1$ shares all but one reaction (the newly added reaction), every parent genotype in this set is adjacent to every other parent genotype. In other words, these genotypes form a clique in $V(n + 1)$ [21].

We next point out that if two genotypes of size $n$ are adjacent, then their corresponding genotypes of size $(n+1)$ form two cliques linked by at least one genotype of size $(n+1)$. The hypothetical example in Figure 2 illustrates this fact. Consider a "universe" of only $N=6$ reactions - $\{R_1, R_2, R_3, R_4, R_5, R_6\}$. The upper half of Figure 2 shows two hypothetical genotypes ($G_1$ in blue and $G_2$ in red) that are viable, adjacent, and contain two reactions each ($n=2$). Genotype $G_1$ comprises reactions $\{R_1, R_2\}$, while the other genotype $G_2$ comprises reactions $\{R_2, R_3\}$. The lower part of the figure shows all genotypes containing three reactions each that can be obtained from adding one reaction to genotypes $G_1$ and $G_2$. Blue genotypes are parents of $G_1$, whereas red genotypes are parents of $G_2$. Note that the red and blue genotypes form two cliques. Among the 7 genotypes of size $n+1$ that are parents of either $G_1$ or $G_2$, one is special, because the two cliques share it. In our example, this is the genotype containing reactions $\{R_1, R_2, R_3\}$. More generally, this shared genotype is the one genotype obtained from a pair of adjacent genotypes $G_1$ and $G_2$ in $V(n)$ by adding the reaction to $G_1$ that it does not share with $G_2$, or vice versa. (There is only one such reaction, because $G_1$ and $G_2$ are adjacent). We note that additional edges connect genotypes in both cliques (Figure 2). Specifically, those edges connect the genotypes derived from adding the same reaction to $G_1$ and $G_2$. There are exactly ($N$-$n$-1) such edges.

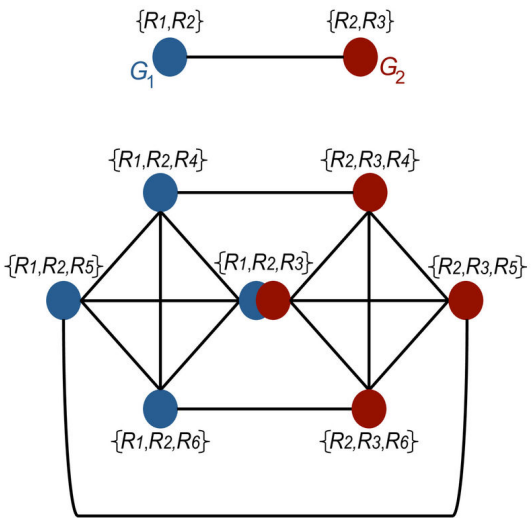



**Figure 2: Connectivity of potential metabolisms can be inferred from parent and child relationship.** The figure uses a hypothetical example of two neighboring metabolisms with three reactions each (upper panel) to illustrate the relationship between the connectedness of genotypes with $n$ reactions ($G(n)$) and their "parents" of $n+1$ reactions that can be obtained from them by adding a single reaction (lower panel). Importantly, if genotypes $G(n)$ form a connected set, then all genotypes $G(n+1)$ obtained by adding one reaction to each of them also form a connected set.

These observations have the following important corollary: If a genotype network containing genotypes of size $n$ is connected, then all genotypes $G(n+1)$ obtained from genotypes of size $n$ are also connected.

So far, our line of reasoning explains connectedness of genotypes that are parents of connected genotypes at a lower size. But some viable genotypes are not parents of any other genotype. These are exactly those genotypes in which elimination of any one reaction abolishes viability. We have called such genotypes *minimal* [10, 12], and note that they do not necessarily correspond to the smallest metabolisms. For example, there are 8 metabolisms that are viable on glucose and that have 24 reactions, all of which are essential (Table 1), but the smallest viable metabolisms on glucose have only 23 reactions. (We explain further below why minimal metabolisms may vary in their number of reactions.). As one increases the size of a metabolism, such "childless" metabolisms could in principle arise at any $n$. Since our preceding argument about connectedness does not apply to them, they need to be identified, and their connectedness to the rest of a genotype network needs to be examined separately, as discussed in the next section.

| **Number of reactions in a metabolism ($n$)** | **Number of components** | **Number of viable metabolisms $V(n)$** | **Number of minimal metabolisms** | **Fraction of minimal metabolisms** |
|---|---|---|---|---|
| 23 | 2 | 3 | 3 | 1 |
| 24 | 2 | 91 | 8 | 0.08791 |
| 25 | 2 | 1333 | 23 | 0.01725 |
| 26 | 2 | 12512 | 14 | 0.00111 |
| 27 | 1 | 84344 | 27 | 0.00032 |
| 28 | 2 | 434238 | 43 | $9.9 \times 10^{-5}$ |
| 29 | 1 | 1773969 | 28 | $1.57 \times 10^{-5}$ |
| 30 | 1 | 5900578 | 15 | $2.54 \times 10^{-5}$ |

**Table 1: The number of minimal metabolisms in genotype networks from the central carbon metabolism** The left-most column shows the number of reactions $n$ in a potential metabolism, the second column from the left shows the number of disconnected components into which the genotype network of these viable metabolisms fragments, the third column shows the number of viable potential metabolisms for each $n$, and the fourth column shows the number of minimal metabolisms. Note that the fraction of minimal metabolisms (column five) decreases as metabolism size $n$ increases.

We identified minimal metabolisms at each size $n$ by deleting every single reaction from each genotype in $V(n)$, and by examining whether the resulting genotype was viable, and thus identifying those genotypes in which no reaction can be deleted. Table 1 shows the number of minimal metabolisms at each $n$, and demonstrates that their proportion among all viable metabolisms ($V(n)$) decreases dramatically with increasing metabolism size $n$. Importantly for the next section, no minimal metabolisms viable on glucose exist above $n = 30$.

In sum, we here observed that if a genotype network is connected at size $n$, the genotype network formed by the parents of its genotypes is also connected. Because minimal metabolisms are not parents of any other metabolisms, they need to be analyzed separately.

**Metabolic genotype networks are connected for all but the smallest metabolisms**

To determine connectedness of genotype networks for metabolisms *V(n)* viable on glucose, we began by analyzing the smallest ($n = 23$–$28$) and largest ($n = 46$–$50$) potential metabolisms. We did so by computing, first, edge lists for each genotype network, and, second, the connectedness of the genotype network, using the graph analysis software igraph [45]. We found that viable metabolisms of size $n = 27$, as well as $n = 46$ to $n = 50$ have only one connected component. In contrast, viable metabolisms of sizes $n = 23$, 24, 25, 26, and 28 are fragmented. They form a genotype network with two components (Table 2).

| **Number of reactions in a metabolism ( *n*)** | **Number of viable potential metabolisms *V*(*n*)** | **Number of components** | **Fraction of viable metabolisms in the largest connected component** |
|---|---|---|---|
| 23 | 3 | 2 | 0.667 |
| 24 | 91 | 2 | 0.637 |
| 25 | 1333 | 2 | 0.997 |
| 26 | 12512 | 2 | 0.992 |
| 27 | 84344 | 1 | 1 |
| 28 | 434238 | 2 | 0.977 |

**Table 2: Fragmentation occurs in central carbon metabolisms close to minimal number of reactions** . For each metabolism size, the table shows the number of viable potential metabolisms, the

number of components, and the fraction of genotypes in the largest component. The genotype network comprising metabolisms of size 23 contains three metabolisms, of which two form one component and the other is isolated from them. For metabolisms of size 24, the two components are almost of the same size, and the larger component contains 63.74 percent of the viable genotypes. For larger metabolisms, the genotype network is largely connected, with more than 97 percent of genotypes belonging to the largest component.

Fragmented genotype networks may decompose into components with different sizes, such that the majority of genotypes belong to the largest component. In this case, most viable genotypes can be reached from each other through a series of small genotypic changes that affect only single reactions each and that leave the phenotype constant. Alternatively, fragmentation of a genotype network may result in components with similar size, which can impede accessibility of many genotypes. Table 2 shows that this is not generally the case. The vertical axis denotes the fraction of genotypes belonging to the largest component of a genotype network, and it shows that the largest components of the genotype networks at size $n = 25–28$ encompass almost all (i.e., more than 99 percent) of the viable genotypes. At $n = 23$, the genotype network has two components that consist of one and two metabolisms. At size $n = 24$ there are 91 viable genotypes, 58 (64 percent) of which belong to the largest component.

We next turn to a more detailed analysis of genotype network fragmentation at the smallest sizes. Figure 3 shows graph representations of genotype networks whose metabolisms have sizes $n = 23$, 24 and 25. Filled circles represent genotypes. Adjacent genotypes are connected by an edge. The size of a circle corresponds to the number of neighbors of the corresponding genotype. Minimal potential metabolisms are shown in red in all three panels. All three potential metabolisms of size 23 are minimal (Figure 3A). Two of them are adjacent metabolisms and form component *A* (left), whereas the remaining isolated metabolism forms component *B* (right). The green and orange circles of Figure 3B show the result of adding one reaction from the remaining pool of 28 reactions ($N$-$n = 51–23 = 28$) to each of the two genotypes in component *A*. Such addition yields a connected component *A'* of 55 metabolisms with 24 reactions (green, Figure 3B). The component consists of two cliques connected to each other by a single connected genotype. Analogous addition of reactions to the single genotype in component *B* of Figure 3A yields a connected

component *B'* with 28 potential metabolisms (orange) of size 24. The total number of genotypes in component *A'* and *B'* is 83. However, there are 91 viable metabolisms with size 24 (Table 1). It turns out that the missing eight metabolisms are minimal (red) and cannot be derived using reaction addition to metabolisms at size 23. Three of them are connected to component *A'* and five of them to component *B'* (Figure 3B). Overall, the number of components at size 24 reflects the number of components at size 23, because these components are derived from the smaller components at size 23. This, however, is no longer true for the genotype network of metabolisms with 25 reactions in Figure 3C. In this panel, genotypes shown in green (components *A"*) and orange (components *B"*) are parents of the green and orange genotypes in components *A'* and *B'* respectively. Notice that these components are now connected, in contrast to their disconnectedness at size 24. What connects them are some of the minimal metabolisms that arose anew at size 24 and 25. There are 23 such child-less minimal genotypes at size 25 (Figure 3C and Table 1). Four of them form a new component labeled *C* (center bottom of Figure 3C).

An analogous analysis of metabolisms up to size 30 can help understand why all larger metabolisms must be connected (Additional files 5 and 6). There are two germane observations. First, at size $n = 30$, there are approximately $5.9 \times 10^6$ metabolisms and all of them fall into a single connected component (Table 1). Second, no minimal metabolisms exist at size 31 and beyond (Table 1). This means that all parent metabolisms at size $(n + 1)$ are derived from child metabolisms at sizes beyond $n = 30$. By our argument in the preceding section, they must therefore form a single connected component (Figure 2).

In sum, we showed that genotype networks formed by different central carbon metabolism variants are connected in metabolic genotype space for all but the smallest viable metabolisms. With few exceptions, wherever fragmentation occurs, more than 99 percent of genotypes belong to the largest component. This high connectivity arises from the parent–child relationships we discussed, as well as from the relatively small number of minimal metabolisms that arise at each *n* (Figure 2 and Table 1).

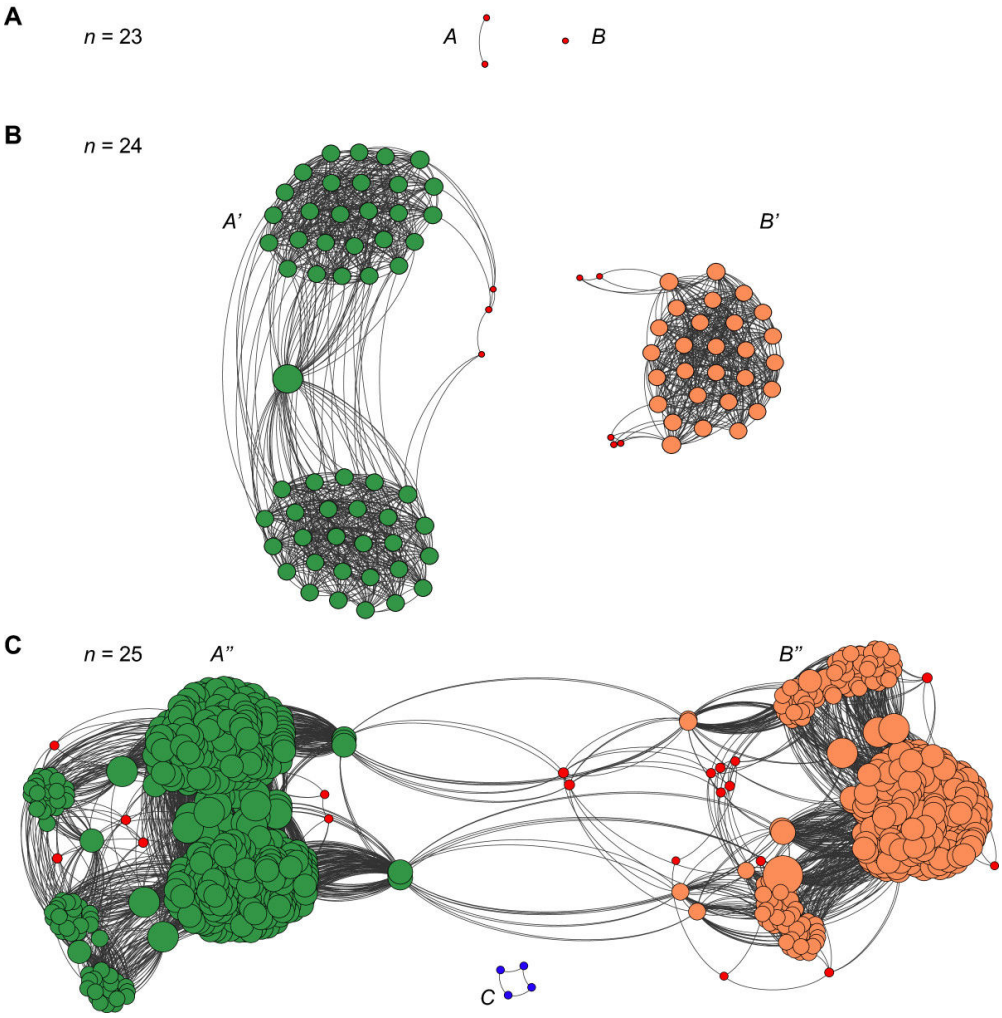



Figure 3 **Organization of metabolic genotype space.** The figure shows the genotype networks of potential metabolisms containing 23, 24 and 25 reactions. Each filled circle corresponds to a genotype. Two genotypes are connected by an edge (curved line) if they are neighbors. Red circles correspond to minimal metabolisms of a given number of reactions *n*. **(A)** The genotype network of size 23 is fragmented, with component *A* containing two adjacent genotypes, while component *B* contains one genotype. **(B)** Structure of the genotype network at size $n = 24$ reactions. Addition of one reaction to the two genotypes in component *A* results in genotypes of size 24 which belong to component *A'* (green), and addition of one reaction to the genotype in component *B* yields genotypes of size 24 which belong to component *B'* (orange). At size $n = 24$ reactions, eight minimal metabolisms (red circles) also arise, of which three genotypes belong to component *A'*, and five to component *B'*. Note that genotypes in components *A'* and *B'* remain disconnected. **(C)** Structure of the genotype network at size $n = 25$ reactions. Adding one reaction to all genotypes in component *A'* yields genotypes (green) in subgraph *A"*, while adding one reaction to all genotypes in component *B'* yields the genotypes (orange) of subgraph *B"*. There are 23 minimal metabolisms of size 25 (red), of which 4 genotypes form a disconnected component *C* (blue, bottom center). Note that genotypes of size 25 in subgraphs *A"* and *B"* are connected either directly or through minimal metabolisms. The size of each circle corresponds to its number of neighboring genotypes, which increases as metabolism size increases. Graphs were drawn using the graph visualization software Gephi [46].

**Essential pathways cause genotype network fragmentation**

Thus far, our analysis focused on broad patterns of genotype network fragmentation. We next discuss the possible mechanistic reasons for such fragmentation. They revolve around different biochemical pathways that are essential for viability among metabolisms in different components. Essential reactions are those whose removal results in a loss of viability (see Methods), and a reaction's essentiality may depend on other reactions present in a metabolism. That is, a reaction can be essential in one potential metabolism, but nonessential in another potential metabolism, because of the presence of alternative metabolic routes [13]. The fraction of metabolisms of a given size in which a reaction is essential is a useful quantifier of the reaction's essentiality, which we have called the reaction's superessentiality index [13]. The concept of (super)essentiality can be extended to entire metabolic pathways, groups of essential reactions that share substrates/products with each other and cannot be replaced without a loss of viability.

We next illustrate with an example how pathway (super) essentiality causes fragmentation of genotype networks, by demonstrating the existence of alternative essential pathways in different network components for metabolisms with 23 and 24 reactions. To identify such pathways, we first computed the superessentiality index of reactions in potential metabolisms of size 23 and 24 each, and did so for all genotypes in each of the two genotype network components (Figure 3A and B) separately (see Methods). We then examined which reactions differ in their superessentiality index between the two components. We found five such reactions, which can be subdivided into groups of two and three reactions, respectively. The first group comprises the reactions catalyzed by transketolase 1 (TKT1) and transaldolase (TALA). They are essential in all metabolisms from network component *A'* (Figure 3), but inessential in all metabolisms belonging to component *B'* . The second group comprises the reactions catalyzed by the enzymes glucose-6-phosphate dehydrogenase (G6PDH), 6-phosphogluconolactonase (PGL), and phosphogluconate dehydrogenase (GND). They are essential in all metabolisms of component *B'*, but inessential in any of the genotypes in component *A'* (Figure 3). Taken together, this means that TKT1 and TALA form a small but essential pathway in the genotypes belonging to component *A'*, while G6PDH, PGL and GND form another essential pathway in genotypes belonging to component *B'*.

These five reactions are part of the pentose phosphate pathway, as shown in Figure 4. The pentose phosphate pathway is required for the synthesis of two biomass precursors, ribose-5-phosphate (r5p) and erythrose-4-phosphate (e4p) (solid squares in Figure 4). The reactions shown in black are essential in potential metabolisms belonging to both genotype network components (Figure 3A and B). In contrast, the essentiality of reactions participating in the two alternative essential pathways (green and orange), which contain the reactions discussed in the preceding paragraph, depends on which of the two components a potential metabolism belongs to. To understand why, we first note that the metabolites glucose-6-phosphate (g6p), fructose-6-phosphate (f6p), and glyceraldehyde-3-phosphate (g3p) are also synthesized by reactions in glycolysis, and thus constitute metabolic inputs to the pentose phosphate pathway for the synthesis of e4p and r5p. Flux balance analysis can be used to show that reactions catalyzed by transketolase 1 (TKT1) and transaldolase (TALA) are required to synthesize sufficient r5p for viability (Table S1 - biomass reaction) upon removal of any one reaction from the orange pathway (G6PDH, PGL, GND), thus rendering the reactions catalyzed by TKT1 and TALA essential. Conversely, removal of any one reaction from the green pathway (TKT1, TALA) leads to a requirement for all reactions in the orange pathway to produce the pathway output. In sum, the genotypes of size 23 and 24 are disconnected because alternative essential pathways exist in them that consist of more than one essential reaction, and because no one reaction in one pathway can replace a reaction in the other pathway. Put differently, loss of any one reaction in one pathway can only be compensated by addition of all reactions of the other pathway. Because metabolisms at size 23 are separated by three swaps, genotype space can be connected at size 25 (subgraphs *A"* and *B"*), that is, after successive addition of two reactions.

In Additional files 5 and 7 (section - Essential pathways cause genotype network fragmentation) we discuss another example, which illustrates that essential and alternative metabolic routes need not contribute to biosynthesis of the same precursors, and may arise in functionally different and unrelated parts of metabolism. These differences notwithstanding, the examples illustrate the mechanistic reason for genotype network fragmentation: It is not possible to interconvert two genotypes in different components by one reaction swap because such interconversion will

inevitably create unviable genotypes in which two alternative essential pathways are incomplete.

As a corollary, the longer such essential alternative pathways are, the greater the number of reactions $m$ that need to be added to non-adjacent viable genotypes $G(n)$, such that viable genotypes $G(n+m)$ become connected.

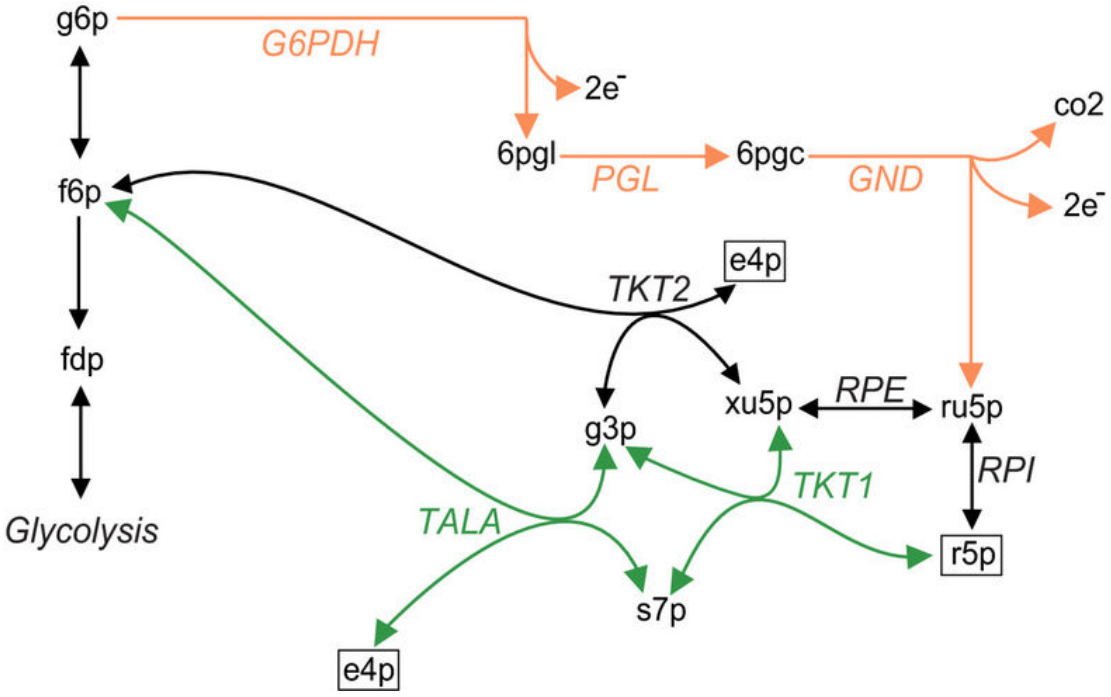



**Figure 4 Essential pathways in pentose phosphate metabolism.** The figure shows the two essential pathways (orange and green) in pentose phosphate metabolism that are necessary for the synthesis of biomass precursors e4p and r5p (in square boxes). Reactions in black are essential regardless of the metabolism in which they occur. Reactions catalyzed by G6PDH, PGL and GND form an essential pathway (orange), while reactions catalyzed by TALA and TKT1 (green) form another essential pathway. If the reactions in orange are absent, the reactions in green become essential and vice versa. This is because removal of TALA and TKT1 requires the synthesis of r5p through the reactions catalyzed by G6PDH, PGL and GND, while removal of these reactions forces the synthesis of r5p through TAL and TKT1. Note that metabolites g6p, f6p and g3p also participate in glycolysis and therefore can be produced there and supplied to the pentose phosphate pathway. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Abbreviations - g6p, D-glucose-6-phosphate; r5p, D-ribose-5-phosphate; e4p, D-erythrose-4-phosphate; f6p, D-fructose-6-phosphate; fdp, fructose-diphosphate; g3p, glyceraldehyde-3-phosphate; G6PDH, glucose-6-phosphate dehydrogenase; PGL, 6-phosphogluconolactonase; GND, 6-phosphogluconate dehydrogenase; RPI, ribose-5-phosphate isomerase; RPE, ribose-5-phosphate 3-epimerase; TKT1, transketolase 1; TALA, transaldolase; TKT2, transketolase 2.

**Metabolisms viable on multiple carbon sources are also mostly connected**

Many organisms are viable on multiple carbon sources, which may impose additional constraints on a metabolism. We wished to find out how severely these constraints affect genotype network connectivity in our analysis of central carbon metabolism. To this end, we analyzed metabolisms that are a subset of our $N = 51$ reactions and that are viable on a total of 10 common carbon sources when each of them is provided as the sole carbon source (see Methods for all these carbon sources). Because glucose is among these 10 carbon sources, metabolisms viable on all 10 carbon sources are also viable on glucose. In other words, the genotype network they form at any specific metabolism size $n$ is a subset of the genotype network of metabolisms viable on glucose. We note that the metabolism comprising all $N = 51$ reactions is viable on all 10 different carbon sources.

We used an approach identical to that described above for glucose to identify potential metabolisms viable on all 10 carbon sources. Their numbers are shown in Figure 1A (blue circles), which shows that, first, no metabolism with fewer than $n = 34$ reactions is viable on all 10 carbon sources, whereas the minimal size is much smaller ($n = 23$) for metabolisms viable on glucose alone (Additional file 3, Additional file 8, Table 1). Second, the number of metabolisms viable on 10 carbon sources is much smaller than the number of metabolisms viable on glucose. It reaches a maximum at $n = 42$ with $2.1 \times 10^5$ metabolisms (Additional file 8), many fewer than for viability on glucose ($2.39 \times 10^8$ metabolisms at $n = 37$). This difference is also highlighted in Figure 1B whose vertical axis represents genotypes viable on glucose (grey) and on all ten different carbon sources (blue) as a fraction of all genotypes. At the minimal size of $n = 34$ reactions, genotypes viable on all 10 different carbon sources comprise approximately one $10^{-8\text{th}}$ of those genotypes viable on glucose of the same size.

This strong constraint on metabolisms viable on multiple carbon sources raises the possibility of genotype network fragmentation. However, we found no evidence for such fragmentation. Because the number of genotypes viable on 10 carbon sources is relatively small, we were able to use standard algorithms to determine their connectedness, which show that genotype networks of all sizes except for $n = 35$ and 36 reactions consist of only one connected component. At size $n = 35$ the genotype

network fragments into three components. The largest of them contains 91.13 percent of viable genotypes. At size $n = 36$, the network fragments into two components, with the larger containing 99.6 percent of genotypes. This implies that one can access any metabolic genotype viable on 10 carbon sources, regardless of its size, from most other viable genotypes through a series of individual reaction changes.

**Study system 2: Genome-scale metabolisms**

We have thus far studied connectedness for potential metabolisms drawn from the reduced reaction set of central carbon metabolism, which comprises a small subset of the more than 1000 reactions in the typical metabolism of a free-living organism. In this section, we focus on the connectedness of larger, genome-scale potential metabolisms. Their reactions come from the known "universe" of possible biochemical reactions, which comprises, at our present state of partial knowledge, already more than 5000 reactions [31, 32]. For any one such metabolism to be viable, we require that it is able to synthesize all 63 essential biomass precursors of *E. coli*[30] – most of which are molecules central to all life, such as nucleotides and amino acids (see Methods) – in a minimal environment containing glucose as the sole carbon source.

Using our binary representation of a metabolic genotype (Additional file 1), the number of possible genome-scale metabolisms is greater than $2^{5000}$, which renders exhaustive analysis of connectivity infeasible. Random sampling of the space using Markov Chain Monte Carlo (MCMC) methods can be very useful [10, 11, 13, 33], but it is not suitable for our purpose, because the MCMC approach samples genotypes from the same component of a genotype network.

We thus use a different sampling approach [10, 12, 33, 47], which starts from a "global" metabolism that comprises all reactions in the known universe (and is viable on glucose). This metabolism has 5906 reactions. Its viable children would form a single connected component, but as one reduces their number of reactions further, the set of viable genotypes $V(n)$ might become disconnected. Figure 5A illustrates this possibility schematically. It shows a funnel-like landscape whose width at a given number of reactions $n$ (vertical axis) indicates the number of viable metabolisms at this $n$. The number of viable metabolisms approaches zero as $n$ approaches the

smallest possible size at which a metabolism can be viable. Starting from the global metabolism, one can randomly select a sequence of reactions for deletion while requiring that each deletion retain viability. Parts of three hypothetical deletion sequences are shown as three trajectories in the panel. Two of them (solid) lead into deep depressions in the funnel, which correspond to disconnected components of a genotype network. More precisely, a metabolism that resides in one such depression cannot be converted into another viable metabolism without changing its number of reactions (the altitude in the landscape), as doing so would require it to traverse the exterior of the funnel. The third trajectory (dotted line) enters such a depression only at a much lower number of reactions. We wanted to know whether such funnels appear in the landscape at moderate *n* (Figure 5A) or only at values of *n* close to the smallest number of reactions permitting viability (Figure 5B).

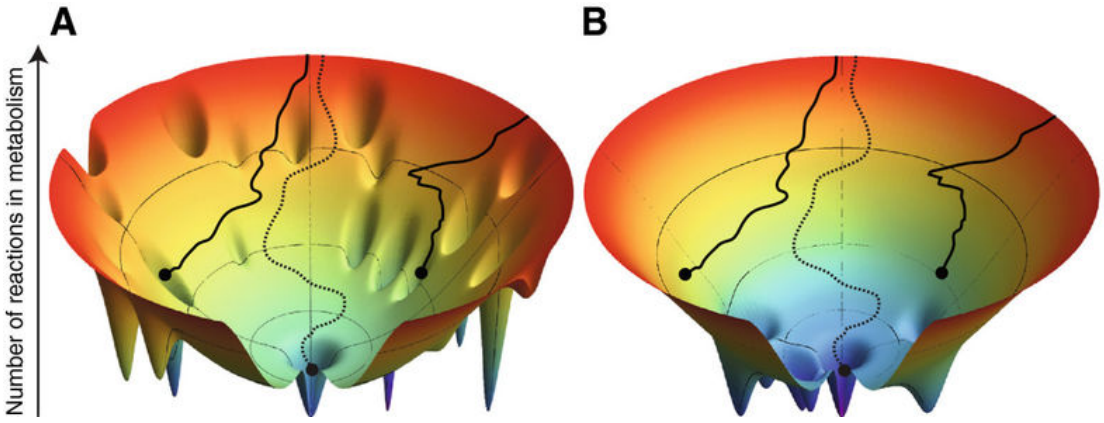



Figure 5

**A spatial schematic of genotype network connectivity at different metabolism sizes** n **.** Each panel shows a funnel-like landscape, where the funnel's width reflects the number of potential metabolisms that are viable at any given number *n* of reactions (altitude in the landscape). The lowest points of the landscape correspond to potential metabolisms whose size are close to the smallest possible *n* needed for viability. Solid and dotted lines denote random sequences of reaction removal from some viable starting metabolism (not shown) that terminate at a particular size *n* and result in some viable metabolism denoted by circles. Depressions in the funnel correspond to disconnected components of the genotype network. Disconnected components in **(A)** arise at higher metabolism sizes than in **(B)**. Two of the trajectories in A (solid lines) terminate in such depressions, that is, in potential metabolisms that are part of a disconnected component. These metabolisms are not interconvertible through viability preserving reaction swaps. The same trajectories in B terminate at a part of the funnel where all viable metabolisms are still connected. The figure was generated using a script from (http://www.oaslab.com/Drawing_funnels.html)

To find out, we derived multiple viable metabolisms with a given size *n* as follows. Starting from the global metabolism, we repeatedly deleted randomly chosen reactions from it, such that each deletion preserved viability, until we had arrived at a minimum metabolism, that is, a metabolism whose number of reactions cannot be reduced further. In doing so, we kept track of the deleted reactions, and the sequence in which they were deleted. Each minimum metabolism created in this way had fewer than 400 reactions (see also below). We used these minimal metabolisms, as well as information about the sequence in which reactions were deleted, to create larger potential metabolisms of varying sizes *n*, each of which corresponds to a specific point in the deletion sequence. We repeated this procedure 500 times, which allowed us to create 500 minimal metabolisms, as well as 500 potential metabolisms of various intermediate sizes.

**Most viable genome-scale metabolisms reside in the same connected component**

If genotype networks were highly fragmented at a given size *n*, then different random deletion sequences would yield potential metabolisms that reside in different components of a genotype network. In this case, it would not be possible to connect two metabolisms that reside in different components of a genotype network through a sequence of reaction swaps, each of which preserves viability. With these observations in mind, we attempted to connect metabolisms in our samples of viable metabolisms of a given size (see Methods). Specifically, for any sample of metabolisms [$G_1$, $G_2$, $G_3$, … , $G_{500}$], we attempted to connect $G_i$ and $G_{i+1}$ ($1 \leq i < 500$) through viability-preserving reaction swaps. We did this for 500 potential metabolisms of size 1400 (similar to that of *E. coli*), 1000, 500, and 400 (above the size of minimal metabolisms, see below). In this way, we were able to show that all 500 potential metabolisms are connected at each of these sizes. Thus, down to a size of $n = 400$ reactions, the genotype network of metabolisms viable on glucose is not highly fragmented, and one component comprises the vast majority or all metabolisms.

Because many free-living microorganisms are viable on multiple carbon sources, we generated 500 additional potential metabolisms through the reaction deletion process just described, but with the additional constraint that they remain viable on ten sole carbon sources (the same ten as used in our analysis of central carbon metabolism).

Specifically, we created again potential metabolisms of size 1400, 1000, 500, 450 and 425 (slightly above the size of minimal metabolisms for viability on 10 carbon sources). We then repeated the procedure that attempts to connect genotypes $G_i$ and $G_{i+1}$ through viability-preserving reaction swaps. In this way, we were able to show that all 500 potential metabolisms are connected at each of these sizes. Thus, down to a size of $n = 425$ reactions, the genotype network of most metabolisms viable on all 10 carbon sources consists of one connected component.

It is possible to make this point more quantitatively and establish a statistical bound on the fraction of potential metabolisms contained in the largest connected component of *V(n)*. Specifically, let us consider the null hypothesis that more than one percent of *V(n)* resides outside this largest component. If this null hypothesis is correct, then the probability $p$ that a randomly drawn viable genotype is *not* on this largest component is greater then $p = 0.01$. Moreover, the probability that some number $M$ of genotypes drawn at random from *V(n)* all fall on the largest connected component would be smaller than $(1\text{-}p)^M$. In our case, $M = 500$ and $(1\text{-}p)^M < 0.99^{500} = 0.0066$. In other words, the results of our sampling allow us to reject the above null hypothesis at a significance level smaller than 1 percent.

**Minimal metabolism size can help explain connectedness**

In the sections on central carbon metabolism we showed that new components disconnected from the remainder of a genotype network can arise as one increases metabolism size, and that they originate from "childless" minimal metabolisms which appear at a given size $n$ that is small compared to the total number of possible reactions. To examine their size for larger metabolic system, we studied the 500 minimal metabolisms that we derived from the sequential random deletion strategy described in the previous section (Figure 6A). Their size ranges from 324 to 391 reactions, with a mean of 352 reactions (standard error = 11.44 reactions) (Figure 6A). Although we cannot absolutely exclude the possibility that minimal metabolisms exist with more than 400 reactions, the fact that all of the minimal metabolisms we found have fewer reactions suggests that the emergence of new genotype network components will be rare above 400 reactions. This observation further supports our assertion that most metabolisms with more than 400 reactions will be part of a single genotype network. It also means that essential alternative

metabolic pathways of more than one reaction that are characteristic for a given connected component exist only for small metabolisms. Alternative pathways for the synthesis of most biomass molecules undoubtedly exist, but most of them can be converted into one another through sequences of single reaction changes that preserve viability.

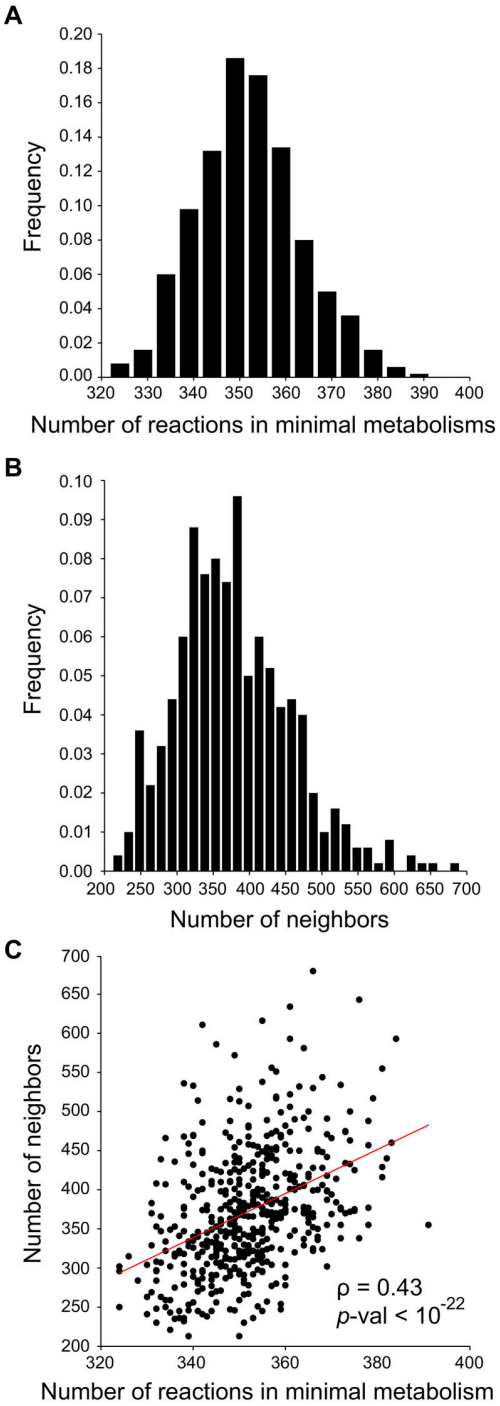



**Figure 6 Minimal metabolisms from the complete universe can have many viable neighbors. (A)** The horizontal axis denotes the size of minimal metabolisms and the vertical axis denotes their frequency. The average minimal metabolism comprises 352 reactions, while the largest minimal metabolism we find has 391 reactions. **(B)** The vertical axis shows the frequency of potential metabolisms with a given number of neighbors (horizontal axis). A minimal metabolism has 372.8 viable neighbors on average. Data in **(C)** show that the number of viable neighbors (vertical axis) is positively correlated with the number of reactions present in a minimal metabolism (horizontal axis). Data in **(A)**, **(B)** and **(C)** are based on 500 minimal metabolisms generated through the random reaction deletion process described in the text.

In a final analysis, we asked whether the minimal networks that our approach identified are isolated in metabolic genotype space $\Omega(n)$, or whether they might themselves form large components. To this end, we simply asked whether these networks have any viable neighbors in $\Omega(n)$, metabolisms that differ by a single reaction swap, which are also viable. The result (Figure 6B) shows that even minimal metabolisms have typically hundreds of neighbors. Specifically, an average minimal metabolism has 372.8 viable neighbors (standard deviation: 79 neighbors). The maximum number of neighbors for a minimal metabolism is 685. Figure 6C shows that larger minimal metabolisms tend to have more neighbors than smaller ones (Spearman's $\rho = 0.43$, $p$-value $< 10^{-22}$). Taken together, this means that minimal metabolisms themselves must form large components and are certainly not isolated. It mirrors the situation in central carbon metabolism, where newly emerging minimal metabolisms at a given size $n$ also form connected components, albeit small ones (Figure 3).

## 5.4. Discussion

To our knowledge, our analysis of all possible $\approx 10^{15}$ metabolisms comprising subsets of reactions in central carbon metabolism is the first exhaustive analysis of a metabolic space this large, even though smaller-scale analyses were carried out before with different goals [48, 49, 50]. Our analysis focused on metabolisms viable on glucose, which are required to synthesize 13 products of central carbon metabolism that are biomass precursors. We found that viable metabolisms could have fewer than half (23) of the maximal number of 51 reactions in central carbon metabolism. Moreover, for metabolisms covering 77 percent of the size of the viable range ($n = 23$–$51$), all ($n = 29$–$51$) or the vast majority of metabolisms of size $n$ form a single connected component (network) in the space of metabolisms.

In genome-scale metabolisms, where exhaustive enumeration is no longer possible, and where we required the synthesis of 63 common biomass molecules for viability, we found viable potential metabolisms with as few as 324 reactions, and for 93.23 percent of the size range of viable metabolisms ($n = 400$–$5906$) the vast majority of potential metabolisms form a single connected component of a genotype network. More specifically, with a probability of greater than 0.99, more than 99 percent of all

viable metabolisms exceeding 400 reactions are part of the same component. We note that it would have been sufficient to perform the sequential reaction deletion procedure needed to arrive at this conclusion for metabolisms of size $n = 400$, and not also for metabolisms of size $n = 400–1400$, as we did. The reason is an elementary observation we made about metabolic genotype space: If a set of viable metabolisms $V(n)$ is connected for some number of reactions $n$, then $V(m)$ must be connected for all $m > n$, provided that no new minimal metabolisms appear at any value of $m$. The largest minimal metabolism we found has $n = 391$ reactions, and while we cannot exclude the existence of minimal metabolisms above $n = 400$ with certainty, such potential metabolisms would be increasingly rare at large $n$. They would create new genotype network components that would comprise a vanishing fraction of the rest of the connected genotype network (even though they might contain many potential metabolisms in absolute numbers).

Figure 5B illustrates schematically the dependence of fragmentation on metabolism size $n$ that we observed. Depressions in the funnel-like landscape whose width reflects the number of viable potential metabolisms correspond to disconnected metabolic networks and appear only at small altitudes (metabolism sizes). That is, the hypothetical landscape of Figure 5B reflects our observations, whereas that of Figure 5A, where disconnected metabolisms appear at much higher reaction numbers does not.

While we study only viability on a carbon source, other metabolic properties such as mutational robustness and access to novel phenotypes are also important [10, 11] and may differ in different components. In such cases, historical contingency may indeed play a role towards the fine-tuning of metabolic properties and would be relevant in a scenario depicted by Figure 5A. However, as fragmentation occurs only at lower metabolism sizes, historical contingency may not constrain the overall evolution of metabolic systems sharply.

With possible exceptions in some marine bacteria [51, 52] metabolisms with sizes as small as $n = 400$ are not usually found in free-living organisms. They occur in (endo)symbionts [53, 54] and (endo)parasites [55, 56], which live in close association with a host organism and are provided nutrients and a constant environment which allows them to shed many enzyme-coding genes [47, 57, 58, 59, 60]. Organisms that

have lived inside a host for a long time experience less of the kinds of evolutionary change – especially horizontal gene transfer – that is powerful in endowing the genomes of free-living organisms with new evolutionary adaptations [57, 59]. In other words, the fragmentation of genotype networks that we see for very small potential metabolisms, and that can constrain their evolution, is of little relevance for the evolution of free-living organisms. Those organisms whose evolution it could constrain the most are already subject to little evolutionary change for ecological reasons.

Our analysis of genotype network fragmentation provides a coarse, statistical view on the organization of genotype space. This view needs to be complemented by a mechanistic perspective that asks what distinguishes the metabolisms that exist in different components of a genotype network? What could prevent evolution from converting them into each other through a series of single viability-preserving reaction changes? The answer lies in alternative metabolic pathways that are essential for the biosynthesis of one or more biomass molecules. Potential metabolisms in one genotype network component have one such pathway, and potential metabolisms in the other component have another such pathway. (In addition, these potential metabolisms may differ in other essential pathways.) At least one of these pathways must comprise more than one reaction, otherwise the two metabolisms could be converted into one another through a single reaction swap. We have provided two examples, one involving the biosynthesis of erythrose-4-phosphate and ribose-5-phosphate through variants of the pentose phosphate pathway, the other concerning the biosynthesis of phosphoenolpyruvate.

For two reasons, such alternative essential pathways are not likely to hamper the evolution of most metabolic systems. First, we observed fragmentation only for relatively small metabolisms, which means that in larger metabolisms, alternative *essential* pathways with more than one reaction do not exist. They can usually be converted into each other by single reaction changes that do not cause a loss of viability. Second, our analysis required that we impose change through reaction swaps – a reaction addition paired with a deletion – that leave reaction numbers constant. However, this is not usually how evolutionary change in a metabolism's reactions occurs. For example, horizontal gene transfer frequently adds more than one

gene and thus more than one reaction to a metabolism [61, 62, 63]. In a metabolism that harbors one of two alternatives for an essential pathway, a horizontal gene transfer event may introduce the genes of the other pathway. After that, the two pathways may coexist, and the first pathway is free to deteriorate through loss of function mutations in its genes. A potential example of co-existing alternative pathways involves the two pathways responsible for synthesizing isopentenyl diphosphate (the 1-deoxy-D-xylulose 5-phosphate pathway and the mevalonate pathway), a molecule that is required for the synthesis of isoprenoids. Some actinomycetes that harbor both pathways in their complete forms may have obtained the responsible genes through horizontal gene transfer [64]. In sum, common forms of genetic change can help bridge different components of a genotype network where such components exist.

The main limitation of our work comes from the enormous computational cost associated with evaluating the viability of many metabolisms. While our sampling approach for genome-scale metabolisms allowed us to circumvent this problem for any one carbon source, it is possible that viability on a broader range of carbon sources (or sources of other chemical elements) might have led to greater genotype network fragmentation. This possibility is suggested by our analysis of central carbon metabolism, where metabolisms viable on 10 carbon sources must have at least 34 reactions, and fragmentation of genotype networks stops at 37 reactions. However, for genome-scale metabolisms, the metabolism sizes at which fragmentation would cease would increase only modestly with each additional carbon source on which viability is required. This is because previous work has shown that viability on every additional carbon source requires on average the addition of only two reactions to a metabolism [65]. For example, viability on ten additional carbon sources would increase the size of minimal metabolisms by only 20 reactions. Because the number of minimal metabolisms that arise *de novo* with increasing metabolic complexity $n$ is closely linked to the metabolism size at which fragmentation occurs, viable genotype networks $V(n)$ would still remain connected over the vast majority of the range of $n$. Indeed, we found that genome-scale metabolisms viable on 10 carbon sources and comprising 425 reactions are connected in genotype space and belong to the same component. Possible exceptions might involve metabolisms viable on hundreds of different carbon sources, but even environmental generalists are typically not viable

on that many. (The generalist *E. coli* is viable on some 50 alternative carbon sources [30]).

Another limitation of our work is that we only considered viability on carbon sources. We cannot exclude the possibility that viability on sources of different chemical elements may lead to different fragmentation patterns. However, it is unlikely that carbon sources are exceptional in this regard. For example, the minimal size of metabolisms viable on different sulfur sources comprises only 90 reactions, and is thus even smaller than that of metabolisms viable on carbon [12]. The reason is that fewer biomass molecules contain sulfur, an observation that also holds for the two other key elements nitrogen and phosphorus.

A further limitation is that we focus on evolutionary constraints caused by the presence or absence of biochemical reactions, rather than on differences in the regulation of existing enzymes or their encoding genes. Such regulatory constraints can influence important metabolic properties such as biomass growth rate [20]. However, they can also be easily broken through regulatory evolution, even on the short time scales of laboratory evolution experiments [19, 20, 66]. Reaction absence is thus a more fundamental constraint, but we note that the exploration of regulatory constraints remains an important task for future work. Moreover, to understand connectedness as a function of reaction numbers, we had to preserve reaction numbers and analyze connectedness through reaction swaps. We note that a reaction swap can be considered as an addition of a reaction, which does not change viability and a reaction deletion that preserves viability. That is, every reaction swap can be broken down into two biologically relevant changes, and thus genotype network connectivity resulting from reaction swaps also holds for single reaction additions and deletions.

Finally, we do not consider one potential cause of genotype network fragmentation: If one required for viability that biomass precursors need to be synthesized at a high rate, then genotype networks may fragment more often than we observe. However, fast biomass synthesis and its main consequence, rapid cell division, are not universally important outside the laboratory environment. For example, a survey of microbial growth rates shows that many microbes have very long generation times in

the wild [67]. Rapid growth thus may not be a biological sensible requirement for viability in many wild organisms.

## 5.5. Methods

**Flux balance analysis**

Flux balance analysis (FBA) is a constraint-based computational method [34, 70] that can predict synthetic abilities and other properties of metabolisms – complex networks of enzyme-catalyzed biochemical reactions. Any one such network can comprise anything from a few dozen reactions, such as central carbon metabolism [70], to the thousands of reactions in a complex genome-scale metabolism. FBA uses information about the stoichiometry of each reaction to predict steady state fluxes for all reactions in a metabolic network. The necessary stoichiometric information is represented as a stoichiometric matrix, $S$, of dimensions $m \times n$, where $m$ denotes the number of metabolites, and $n$ denotes the number of reactions in a metabolism [34, 70]. FBA assumes that the concentrations of intracellular metabolites are in a steady state, which allows one to impose the constraint of mass conservation on them. This constraint can be written as $Sv = 0$, where $v$ denotes a vector of metabolic fluxes through each reaction in a metabolism. The above equation has a large space of possible solutions, but not all of these solutions may be of biological interest. To restrict this space to fluxes of interest, FBA uses linear programming to maximize a biologically relevant quantity in the form of a linear objective function $Z$[70]. Specifically, the linear programming formulation of an FBA problem can be expressed as

$$\max Z = \max\{c\mathrm{T}v \,||\, Sv = 0, a \le v \le b\}$$

The vector $c$ contains the set of scalar coefficients that represent the maximization criterion. The individual entries of vectors $a$ and $b$, respectively, contain the minimal and maximally possible fluxes for each reaction in $v$. Irreversible reactions can only have fluxes with positive signs, whereas irreversible reactions can have fluxes of both signs.

We are here interested in predicting whether a metabolism can sustain life in a given spectrum of environments, that is, whether it can synthesize all necessary small biomass molecules (biomass precursors) required for survival and growth. For our

analysis of central carbon metabolism, there are 13 such essential precursors (Additional file 2 and [35]). For our analysis of genome scale metabolisms, we use all 63 [30] biomass precursors of *E. coli*, because most of them would be required in any free-living organism. They include 20 proteinaceous amino acids, DNA and RNA nucleotide precursors, lipids and cofactors. We use these biomass precursors to define the objective function and the vector *c*. We employed the package CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve linear programming problems. The computer program required for FBA is available in Additional file 9.

**Growth environments**

Along with the biomass composition and stoichiometric information about a metabolic network, computational predictions of viability require information about the chemical environments that contain the nutrients needed to synthesize biomass precursors. In our analysis of central carbon metabolism, we consider a minimal aerobic growth environment composed of a sole carbon source, along with ammonium as a nitrogen source, inorganic phosphate as a source of phosphorus, as well as oxygen, protons, and water. When studying the viability of metabolisms on different carbon sources, we vary the carbon source while keeping all the other nutrients constant. When we say a particular metabolism is viable on 10 carbon sources, we mean that it can synthesize all biomass precursors when each of these carbon sources is provided as the sole carbon source in a minimal medium. The ten carbon sources we consider are D-glucose, acetate, pyruvate, D-lactate, D-fructose, alpha-ketoglutarate, fumarate, malate, succinate and glutamate.

Our analysis of genome-scale metabolisms requires a minimal environment with more nutrients, i.e., a sole carbon source, ammonium, inorganic phosphate, sulphate, sodium, potassium, cobalt, iron (Fe2+ and Fe3+), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc [30]. For our analysis of genome-scale metabolisms viable on 10 carbon sources, we used the 10 carbon sources from the preceding paragraph.

**The reactions used in the analysis of central carbon metabolism**

We use a global set of reactions in central carbon metabolism, which is based on a published reconstruction of *E. coli* central carbon metabolism [35]. From the published reconstruction [35], we deleted four reactions involved in ethanol synthesis, metabolism and transport. We also grouped the reactions catalyzed by aconitase A and aconitase B into one reaction. We did this mainly to reduce the size of the set of reactions, in order to render the exploration of all variant metabolisms derived from it feasible. The final reaction set consists of $N = 51$ intracellular reactions, and we analyzed the viability of metabolisms comprising all possible $2^{51}$ subsets of this set. The reconstruction in [35] also involves 20 transport reactions, which are necessary to import nutrients or excrete waste products, and which we assume to be present in all metabolisms we studied.

**The known reaction "universe" and the global metabolism**

We refer to the known universe of biochemical reactions as the set of reactions known to occur in some organism based on currently available biochemical knowledge. To arrive at this set, we curated data from the LIGAND database [31, 32] of the Kyoto Encyclopedia of Genes and Genomes [71], which is divided into two smaller databases, the REACTION database and the COMPOUND database. These two databases together provide information about metabolic reactions, participating chemical compounds, and associated stoichiometric information. As described previously [10, 11, 13, 33], we curated reactions from these databases by excluding reactions involving polymer metabolites of unspecified numbers of monomers; general polymerization reactions with uncertain stoichiometry; reactions involving glycans, owing to their complex structure; reactions with unbalanced stoichiometry; and reactions involving complex metabolites without detailed structural information [71]. After curation of these reactions, we added to them all non-redundant reactions from the published *E. coli* metabolic model (*i* AF1260), which comprises 1,397 non-transport reactions [30]. At the end of this procedure, we had arrived at a set of 5,906 non-transport reactions and 5,030 metabolites. We converted this set into what we call a global metabolism by including all *E. coli* transport reactions in this set [30]. Unsurprisingly, the global metabolism can synthesize all biomass precursors of *E. coli* from any of the carbon sources we consider here. We note that this metabolism

may not be biologically realizable, for example, because it may contain thermodynamically infeasible pathways. However, we merely use it as a starting point to create smaller and ultimately minimal metabolisms through the sequential reaction deletion process described below.

**Genotypes, phenotypes and viability**

The genes encoding the enzymes that catalyze a metabolism's reactions constitute the metabolic genotype of an organism. For our purpose, a more compact representation of a metabolic genotype is useful, which represents this genotype as a binary vector whose *i*-th entry corresponds to the *i*-th reaction in some global set or universe of biochemical reactions. This entry will be equal to one if an organism's genome encodes an enzyme capable of catalyzing this reaction, and zero otherwise (Additional file 1). The genotype space of all possible metabolisms comprises $2^N$ metabolisms, where $N$ is the total number of known or considered chemical reactions ($N = 51$ for our analysis of central carbon metabolism, and $N = 5906$ for our analysis of genome-scale metabolisms). Any one organism's metabolic genotype can be thought of as a point in this space. Genotypes (metabolisms) viable in a given chemical environment are those that can synthesize all biomass precursors from nutrients in this environment. Specifically, a metabolism is considered viable on a carbon source if its biomass synthesis rate is more than one percent of the biomass synthesis rate of the central carbon metabolism ($N = 51$) on that carbon source. We note that many of the metabolisms we study here, may not be realized in extant organisms. We thus refer to these metabolisms as potential metabolisms.

**Essential and nonessential reactions**

We define a reaction as essential for viability if its elimination abolishes viability in a given chemical environment. To identify all such essential reactions in a given metabolism, we eliminated each reaction and used FBA to assess whether non-zero biomass growth flux was still achievable. For our analysis of viability on 10 different (sole) carbon sources, we defined a reaction as essential if its elimination abolishes viability on *at least one* of the 10 carbon sources. The computer program required for computing essential and non-essential reactions is available in Additional file 9.

**Identification of viable central carbon metabolisms**

To identify all viable metabolisms by exhaustive enumeration of viability for all $2^{51}$ ($10^{15}$) possible metabolisms in central carbon metabolism would be infeasible. Fortunately, such brute-force enumeration is also not necessary, for two reasons. The first originates from the notion of "environment-general superessential reactions" [13]. These are reactions whose elimination abolishes viability in each of the 10 carbon sources used here. To find such reactions, we converted the universe of central carbon metabolism into a format amenable to FBA analysis, as described earlier in this section. We then deleted each reaction and determined viability on each of the 10 carbon sources. We found six reactions (Additional file 2, in red) that were necessary for biomass synthesis on each source. Any viable central carbon metabolism would require all six reactions, which reduces the number of metabolisms whose viability needs to be evaluated from $2^{51}$ to $2^{(51-6)} = 2^{45} (\approx 10^{13})$.

The second reason derives from a simple observation that reduces the number of genotypes whose viability needs to be determined even more dramatically: Removal of a reaction from an unviable metabolism cannot result in a viable metabolism. This means that among all metabolisms with $n$-1 reactions, we need to evaluate only the viability of those that are derived from viable potential metabolisms with $n$ reactions through removal of one reaction. We incorporated this idea into an algorithm that allowed us to enumerate all viable genotypes [43].

**Sampling of viable genome-scale metabolisms**

To sample large (genome-scale) metabolisms, we started from the global metabolism of 5,906 reactions and deleted (eliminated) from it a sequence of randomly chosen reactions, while requiring that each such deletion preserves viability. Specifically, we chose a metabolic reaction at random and equiprobably among all reactions, deleted it, and used FBA to determine viability of the resulting metabolism. If the metabolism was viable, we accepted the deletion. Otherwise we randomly choose another reaction for deletion, and so on, until we found one whose deletion left the resulting metabolism viable. We also kept a count of the number of successive attempted deletions that resulted in a non-viable metabolism. This count was reset to zero if the deletion of a randomly chosen reaction was successful. Once that count reached 1000,

that is 1000 successive attempts at reaction deletion abolished viability, we considered the metabolism a good candidate for a minimal metabolism. To confirm minimality, we deleted each reaction in this metabolism, and if every such deletion resulted in non-viability, we declared the metabolism to be minimal. The computer program required for generating viable potential metabolisms by random reaction deletion is available in Additional file 9.

**Identification of viability-preserving paths connecting viable genotypes $G_1$ and $G_2$ at arbitrary size n**

To find out whether two genotypes $G_1$ and $G_2$ can be connected to one another through viability-preserving reaction swaps, we used the following heuristic approach. It does not rely on reaction swaps of arbitrary reactions, which we found to be too inefficient, but takes advantage of existing reactions in the two genotypes to accelerate the process. It defines a "walker" genotype $G_1$ and alters it through multiple random steps (reaction swaps) that approach the other, "target" genotype $G_2$. Before starting this walk, we established two lists of reactions $L_1$ and $L_2$. $L_1$ contained all reactions in $G_1$ that were not contained in $G_2$. In this list we placed reactions non-essential in $G_1$ first (and in random order), followed by reactions essential in $G_1$ (also in random order). Conversely, $L_2$ consisted of arbitrarily ordered essential reactions in $G_2$, followed by arbitrarily ordered reactions nonessential in $G_2$.

Each step in the random walk consisted of two parts, i.e., (i) adding to $G_1$ a reaction from $L_2$ (i.e., a reaction essential in $G_2$), and (ii) deleting from $G_1$ a reaction listed in $L_1$. Subsequent steps used subsequent reactions in each list for addition and deletion.

As this walk through genotype space progressed, we continued adding reactions to $G_1$ until all essential reactions from $G_2$ in list $L_2$ had been added to $G_1$, and continued from there on to adding nonessential reactions from $L_2$. During part (ii) of any given step, if none of the remaining reactions in the list could be deleted from walker $G_1$ without losing viability, we reverted the last reaction addition, and chose instead a reaction at random from the universe as a candidate for addition. Before adding it, we ensured that the chosen reaction shared all of its substrates and products with other reactions in the random walker. If the product of the reaction was not shared with

another reaction, we checked if it could be secreted by a transport reaction. A candidate reaction that did not fulfill both criteria would be disconnected from the rest of metabolism, could therefore not possibly contribute to viability [33], and we discarded it, choosing another candidate, and so on, until we had found one that fulfilled both criteria. We then determined if, after the addition of this reaction, some reaction in the list could be deleted from the random walker. If so, we accepted the resulting swap, otherwise we tried another addition, and so on, until we had found an acceptable swap.

The two parts of each step ensure, first, that essential reactions from the target are preferentially added to the walker, thus increasing the likelihood of adding "useful" reactions to $G_1$, perhaps from one of several alternative metabolic pathways. Second, they reduce the chances of yielding an unviable genotype after a reaction deletion. However, the probability that the deletion of a reaction from walker $G_1$ can render it unviable increases with the number of reaction swaps, because past steps may have rendered previously nonessential reactions essential. We therefore also needed to use FBA after each deletion to ensure that $G_1$ retained viability after a reaction deletion.

We continued this guided random walk for as many swaps as needed to reach the target $G_2$, or until we had performed 5000 attempted swaps. In the latter case, we declared $G_1$ and $G_2$ disconnected. We note that this is no proof of disconnectedness, as some path may exist that this procedure cannot find. However, in practice, all our attempts to connect genotype pairs in this way were successful.

The computer program required for checking connectedness between a pair of metabolic genotypes using the above procedure is available as Additional file 9.

**Identification of a metabolism's viable neighbors**

Two metabolisms are adjacent or neighbors of each other with respect to a reaction swap if they differ by one such swap. If the focal metabolism contains $n$ reactions, then there are $N$-$n$ reactions that are not part of the focal metabolism, where $N$ is the total number of reactions a metabolism could possibly have. One can thus obtain a neighbor of the focal metabolism by deleting one of its $n$ reactions and simultaneously adding one of the $N$-$n$ reaction from the universe of reactions. Any metabolism therefore has $n \times (N\text{-}n)$ possible neighbors. To identify the viable

neighbors of a minimal metabolism, we generated all possible $n \times (N\text{-}n)$ neighbors, and used FBA to determine their viability on glucose. (We also note that any minimal metabolism trivially has zero viable neighbors with respect to reaction deletion, and $N\text{-}n$ viable neighbors with respect to reaction additions).

The computer program required for computing the viable neighbors of a metabolic genotype is available as Additional file 9. We used MATLAB (Mathworks Inc.) for all numerical analysis. Genotype space visualization was generated using the script available at (http://www.oaslab.com/Drawing_funnels.html).

## 5.6. References


1. Schuster P, Fontana W, Stadler PF, Hofacker IL: From sequences to shapes and back: a case study in RNA secondary structures. Proceedings Biological sciences/The Royal Society. 1994, 255: 279-284. 10.1098/rspb.1994.0040.
2. Reidys C, Stadler PF, Schuster P: Generic properties of combinatory maps: neutral networks of RNA secondary structures. Bull Math Biol. 1997, 59: 339-397. 10.1007/BF02462007.
3. Ancel LW, Fontana W: Plasticity, evolvability, and modularity in RNA. J Exp Zool. 2000, 288: 242-283. 10.1002/1097-010X(20001015)288:3<242::AID-JEZ5>3.0.CO;2-O.
4. Wagner A: Robustness and evolvability: a paradox resolved. Proceedings Biological sciences. 2008, 275: 91-100. 10.1098/rspb.2007.1137.
5. Dill KA, Ozkan SB, Shell MS, Weikl TR: The protein folding problem. Annu Rev Biophys. 2008, 37: 289-316. 10.1146/annurev.biophys.37.092707.153558.
6. Honeycutt JD, Thirumalai D: The nature of folded states of globular proteins. Biopolymers. 1992, 32: 695-709. 10.1002/bip.360320610.
7. Ciliberti S, Martin OC, Wagner A: Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput Biol. 2007, 3: e15-10.1371/journal.pcbi.0030015.
8. Payne JL, Wagner A: Constraint and contingency in multifunctional gene regulatory circuits. PLoS Comput Biol. 2013, 9: e1003071-10.1371/journal.pcbi.1003071.
9. Samal A: The regulatory network of *E. coli* metabolism as a Boolean dynamical system exhibits both homeostasis and flexibility of response. BMC Syst Biol. 2008, 2: 21-10.1186/1752-0509-2-21.
10. Samal A, Matias Rodrigues JF, Jost J, Martin OC, Wagner A: Genotype networks in metabolic reaction spaces. BMC Syst Biol. 2010, 4: 30-10.1186/1752-0509-4-30.
11. Matias Rodrigues JF, Wagner A: Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol. 2009, 5: e1000613-10.1371/journal.pcbi.1000613.

12. Matias Rodrigues JF, Wagner A: Genotype networks, innovation, and robustness in sulfur metabolism. BMC Syst Biol. 2011, 5: 39-10.1186/1752-0509-5-39.

13. Barve A, Rodrigues JFM, Wagner A: Superessential reactions in metabolic networks. Proc Natl Acad Sci U S A. 2012, 109: E1121-E1130. 10.1073/pnas.1113065109.

14. Jörg T, Martin OC, Wagner A: Neutral network sizes of biological RNA molecules can be computed and are not atypically small. BMC Bioinforma. 2008, 9: 464-10.1186/1471-2105-9-464.

15. Reidhaar-Olson JF, Sauer RT: Functionally acceptable substitutions in two alpha-helical regions of lambda repressor. Proteins. 1990, 7: 306-316. 10.1002/prot.340070403.

16. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H: Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006, 2: 2006.0008

17. Segrè D, Vitkup D, Church GM: Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A. 2002, 99: 15112-15117. 10.1073/pnas.232349399.

18. Schaper S, Johnston IG, Louis AA: Epistasis can lead to fragmented neutral spaces and contingency in evolution. Proceedings Biological sciences/The Royal Society. 2012, 279: 1777-1783. 10.1098/rspb.2011.2183.

19. Fong SS, Joyce AR, Palsson BØ: Parallel adaptive evolution cultures of Escherichia coli lead to convergent growth phenotypes with different gene expression states. Genome Res. 2005, 15: 1365-1372. 10.1101/gr.3832305.

20. Ibarra RU, Edwards JS, Palsson BO: Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002, 420: 186-189. 10.1038/nature01149.

21. Newman M: Networks: an Introduction. 2010, Oxford;New York: Oxford University Press

22. Wagner A: The Origins of Evolutionary Innovations. 2011, USA: Oxford University Press.

23. Ciliberti S, Martin OC, Wagner A: Innovation and robustness in complex regulatory gene networks. Proc Natl Acad Sci U S A. 2007, 104: 13591-13596. 10.1073/pnas.0705396104.

24. Huynen MA: Exploring phenotype space through neutral evolution. J Mol Evol. 1996, 43: 165-169. 10.1007/BF02338823.

25. Fontana W, Schuster P: Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol. 1998, 194: 491-515. 10.1006/jtbi.1998.0771.

26. Bornberg-Bauer E: How are model protein structures distributed in sequence space? Biophys J. 1997, 73: 2393-2403. 10.1016/S0006-3495(97)78268-7.

27. Aguirre J, Buldú JM, Stich M, Manrubia SC: Topological structure of the space of phenotypes: the case of RNA neutral networks. PloS one. 2011, 6: e26324-10.1371/journal.pone.0026324.

28. Boldhaus G, Klemm K: Regulatory networks and connected components of the neutral space. Eur Phys J B. 2010, 77: 233-237. 10.1140/epjb/e2010-00176-4.

29. Neidhardt F, Ingraham J: Escherichia Coli and Salmonella Typhimurium: Cellular and Molecular Biology. Washington, DC: ASM Press, 13-16. 1

30. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 2007, 3: 121

31. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M: LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 2002, 30: 402-404. 10.1093/nar/30.1.402.

32. Goto S, Nishioka T, Kanehisa M: LIGAND: chemical database of enzyme reactions. Nucleic Acids Res. 2000, 28: 380-382. 10.1093/nar/28.1.380.

33. Barve A, Wagner A: A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature. 2013, 500: 203-206. 10.1038/nature12301.

34. Price ND, Reed JL, Palsson BØ: Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol. 2004, 2: 886-897. 10.1038/nrmicro1023.

35. Orth JD, Fleming RMT, Palsson BØ: Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal - Escherichia coli and Salmonella Cellular and Molecular Biology. Edited by: Karp PD. Washington DC: ASM Press, 10.2.1.

36. Meléndez-Hevia E, Waddell TG, Heinrich R, Montero F: Theoretical approaches to the evolutionary optimization of glycolysis–chemical analysis. Eur J Biochem. 1997, 244: 527-543. 10.1111/j.1432-1033.1997.t01-1-00527.x.

37. Entner N, Doudorrof M: Glucose and gluconic acid oxidation of Pseudomonas saccharophila. J Biol Chem. 1952, 196: 853-862.

38. Bar-Even A, Flamholz A, Noor E, Milo R: Rethinking glycolysis: on the biochemical logic of metabolic pathways. Nat Chem Biol. 2012, 8: 509-517. 10.1038/nchembio.971.

39. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R: Glycolytic strategy as a tradeoff between energy yield and protein cost. Proc Natl Acad Sci U S A. 2013, 110: 10039-10044. 10.1073/pnas.1215283110.

40. Romano AH, Conway T: Evolution of carbohydrate metabolic pathways. Res Microbiol. 1996, 147: 448-455. 10.1016/0923-2508(96)83998-2.

41. Huynen MA, Dandekar T, Bork P: Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol. 1999, 7: 281-291. 10.1016/S0966-842X(99)01539-5.

42. Noor E, Eden E, Milo R, Alon U: Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell. 2010, 39: 809-820. 10.1016/j.molcel.2010.08.031.

43. Hosseini S-R: Exhaustive Genotype-Phenotype Mapping in Metabolic Genotype Space. 2013, Zurich, Switzerland: Swiss Federal Institute of TechnologyGoogle

44. Hopcroft J, Tarjan R: Algorithm 447: efficient algorithms for graph manipulation. Commun ACM. 1973, 16: 372-378. 10.1145/362248.362272.

45. Csardi G, Nepusz T: The igraph software package for complex network research. InterJournal Complex Systems. 2006, 1695: 1695

46. Bastian M, Heymann S, Jacomy M: Gephi: an open source software for exploring and manipulating networks. ICWSM. 2009, 2: 361-362.

47. Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD: Chance and necessity in the evolution of minimal metabolic networks. Nature. 2006, 440: 667-670. 10.1038/nature04568.

48. Bar-Even A, Noor E, Lewis NE, Milo R: Design and analysis of synthetic carbon fixation pathways. Proc Natl Acad Sci U S A. 2010, 107: 8889-8894. 10.1073/pnas.0907176107.

49. Meléndez-Hevia E, Waddell TG, Cascante M: The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. J Mol Evol. 1996, 43: 293-303. 10.1007/BF02338838.

50. Mittenthal JE, Clarke B, Waddell TG, Fawcett G: A new method for assembling metabolic networks, with application to the Krebs citric acid cycle. J Theor Biol. 2001, 208: 361-382. 10.1006/jtbi.2000.2225.

51. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ: Genome streamlining in a cosmopolitan oceanic bacterium. Science (New York, NY). 2005, 309: 1242-1245. 10.1126/science.1114057.

52. Dufresne A, Garczarek L, Partensky F: Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 2005, 6: R14-10.1186/gb-2005-6-2-r14.

53. Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S: Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. Nat Genet. 2002, 32: 402-407. 10.1038/ng986.

54. Pérez-Brocal V, Gil R, Ramos S, Lamelas A, Postigo M, Michelena JM, Silva FJ, Moya A, Latorre A: A small microbial genome: the end of a long symbiotic relationship?. Science (New York, NY). 2006, 314: 312-313. 10.1126/science.1130441.

55. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, Weidman JF, Small KV, Sandusky M, Fuhrmann J, Nguyen D, Utterback TR, Saudek DM, Phillips CA, Merrick JM, Tomb JF, Dougherty BA, Bott KF, Hu PC, Lucier TS, Peterson SN, Smith HO, Hutchison CA, Venter JC: The minimal gene complement of Mycoplasma genitalium. Science (New York, NY). 1995, 270: 397-403. 10.1126/science.270.5235.397.

56. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM: Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science (New York, NY). 1995, 269: 496-512. 10.1126/science.7542800.

57. Moran N, Wernegreen J: Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends Ecol Evol. 2000, 15: 321-326. 10.1016/S0169-5347(00)01902-9.

58. McCutcheon JP, Moran NA: Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. Proc Natl Acad Sci U S A. 2007, 104: 19392-19397. 10.1073/pnas.0708855104.

59. Kikuchi Y, Hosokawa T, Nikoh N, Meng X-Y, Kamagata Y, Fukatsu T: Host-symbiont co-speciation and reductive genome evolution in gut symbiotic bacteria of acanthosomatid stinkbugs. BMC Biol. 2009, 7: 2-10.1186/1741-7007-7-2.

60. Thomas GH, Zucker J, Macdonald SJ, Sorokin A, Goryanin I, Douglas AE: A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. BMC Syst Biol. 2009, 3: 24-10.1186/1752-0509-3-24.

61. Pál C, Papp B, Lercher MJ: Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet. 2005, 37: 1372-1375. 10.1038/ng1686.

62. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV: Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. Genome Biol. 2003, 4: R55-10.1186/gb-2003-4-9-r55.

63. Lawrence JG, Roth JR: Evolution of coenzyme B12 synthesis among enteric bacteria: evidence for loss and reacquisition of a multigene complex. Genetics. 1996, 142: 11-24.

64. Boucher Y, Doolittle WF: The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. Mol Microbiol. 2000, 37: 703-716. 10.1046/j.1365-2958.2000.02004.x.

65. Bilgin T, Wagner A: Design constraints on a synthetic metabolism. PloS one. 2012, 7: e39903-10.1371/journal.pone.0039903.

66. Fong SS, Marciniak JY, Palsson BØO: Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J Bacteriol. 2003, 185: 6400-6408. 10.1128/JB.185.21.6400-6408.2003.

67. Vieira-Silva S, Rocha EPC: The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 2010, 6: e1000808-10.1371/journal.pgen.1000808.

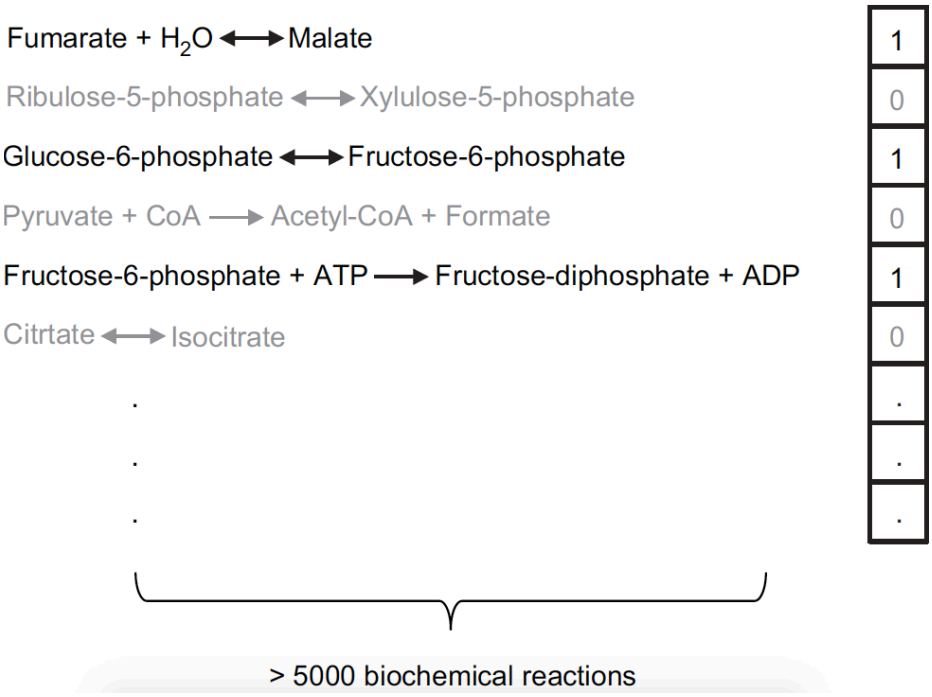
68. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO: Network context and selection in the evolution to enzyme specificity. Science (New York, NY). 2012, 337: 1101-1104. 10.1126/science.1216861.

69. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD: Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol. 2010, 6: 436

70. Kauffman KJ, Prakash P, Edwards JS: Advances in flux balance analysis. Curr Opin Biotechnol. 2003, 14: 491-496. 10.1016/j.copbio.2003.08.001.

71. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010, 38 (Database issue): D355-D360.

## 5.7. Supplementary Information




**Additional file 1: Representation of a genotype vector.** Any genotype encoding $n$ reactions ($n \leq N$) can be represented as a binary vector of length $N$, with $n$ entries equal to one and all others equal to zero. The reactions that are present in the above hypothetical genotype are shown in black and the reactions that are absent are shown in grey. (PDF 10 KB)

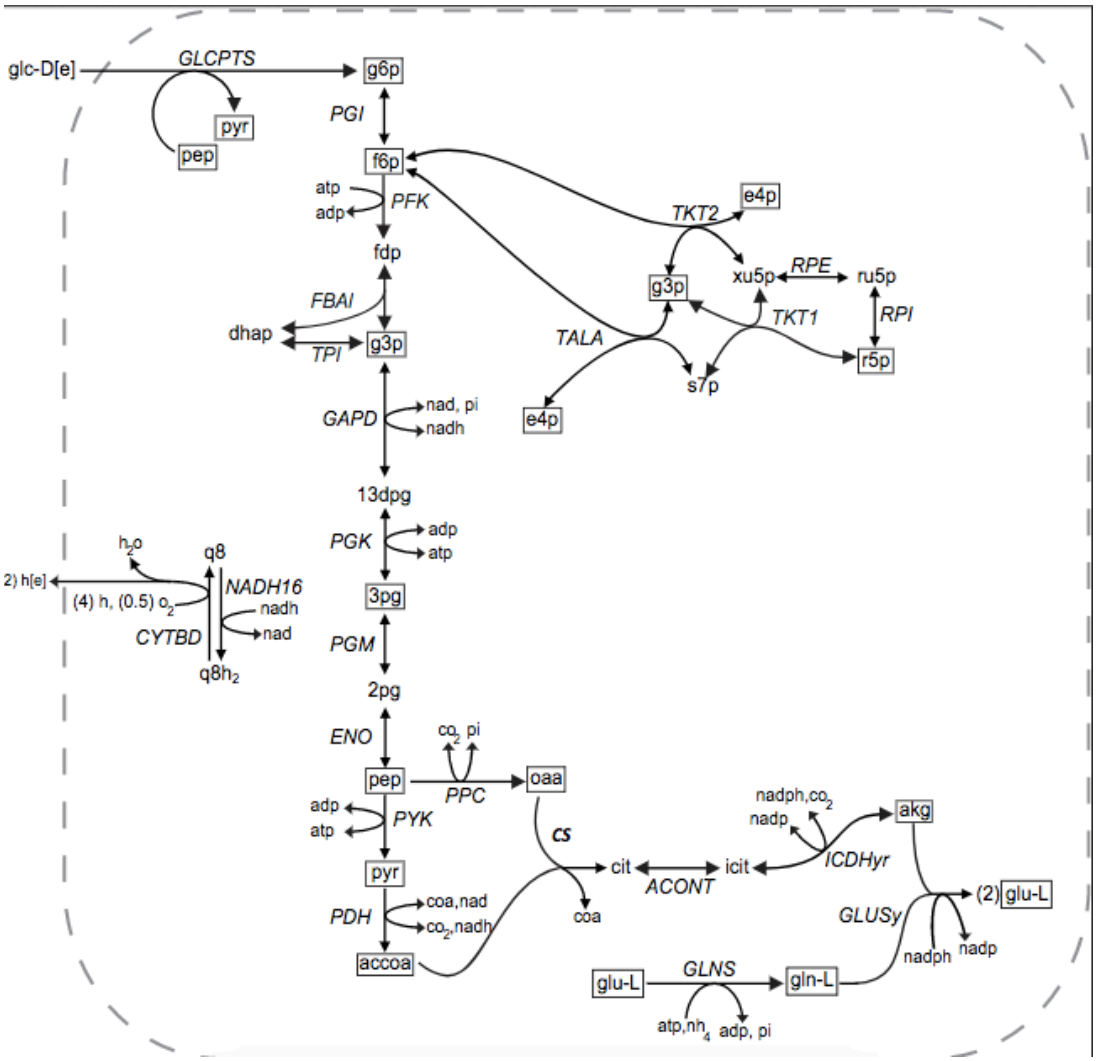
| Abbreviation | Metabolite full name |
|---|---|
| 13dpg | 3-Phospho-D-glyceroyl phosphate |
| 2pg | D-Glycerate 2-phosphate |
| 3pg | 3-Phospho-D-glycerate |
| 6pgc | 6-Phospho-D-gluconate |
| 6pgl | 6-phospho-D-glucono-1,5-lactone |
| ac | Acetate |
| ac[e] | Acetate (extracellular) |
| acald | Acetaldehyde |
| acald[e] | Acetaldehyde (extracellular) |
| accoa | Acetyl-CoA |
| actp | Acetyl phosphate |
| adp | ADP |
| akg | 2-Oxoglutarate |
| akg[e] | 2-Oxoglutarate (extracellular) |
| amp | AMP |
| atp | ATP |
| cit | Citrate |
| co2 | CO2 |
| co2[e] | CO2 (extracellular) |
| coa | Coenzyme A |
| dhap | Dihydroxyacetone phosphate |
| e4p | D-Erythrose 4-phosphate |
| f6p | D-Fructose 6-phosphate |
| fdp | D-Fructose 1,6-bisphosphate |
| for | Formate |
| for[e] | Formate (extracellular) |
| fru[e] | D-Fructose (extracellular) |
| fum | Fumarate |
| fum[e] | Fumarate (extracellular) |
| g3p | Glyceraldehyde 3-phosphate |
| g6p | D-Glucose 6-phosphate |
| glc-D[e] | D-Glucose (extracellular) |
| gln-L | L-Glutamine |
| gln-L[e] | L-Glutamine (extracellular) |
| glu-L | L-Glutamate |
| glu-L[e] | L-Glutamate (extracellular) |
| glu-L[e] | L-Glutamate (extracellular) |
| glx | Glyoxylate |
| h | H+ |
| h[e] | H+ (extracellular) |
| h2o | H2O |
| h2o[e] | H2O (extracellular) |
| icit | Isocitrate |
| lac-D | D-Lactate |
| lac-D[e] | D-Lactate (extracellular) |
| mal-L | L-Malate |
| mal-L[e] | L-Malate (extracellular) |
| nad | Nicotinamide adenine dinucleotide |
| nadh | Nicotinamide adenine dinucleotide (reduced) |
| nadp | Nicotinamide adenine dinucleotide phosphate |
| nadph | Nicotinamide adenine dinucleotide phosphate (reduced) |
| nh4 | Ammonium |
| nh4[e] | Ammonium (extracellular) |
| o2 | O2 |
| o2[e] | O2 (extracellular) |
| oaa | Oxaloacetate |

| | |
|---|---|
| pep | Phosphoenolpyruvate |
| pi | Phosphate |
| pi[e] | Phosphate (extracellular) |
| pyr | Pyruvate |
| pyr[e] | Pyruvate (extracellular) |
| q8 | Ubiquinone-8 |
| q8h2 | Ubiquinol-8 |
| r5p | alpha-D-Ribose 5-phosphate |
| ru5p-D | D-Ribulose 5-phosphate |
| s7p | Sedoheptulose 7-phosphate |
| succ | Succinate |
| succ[e] | Succinate (extracellular) |
| succoa | Succinyl-CoA |
| xu5p-D | D-Xylulose 5-phosphate |

**Additional file 2: Metabolites in central carbon metabolism.** Metabolites abbreviations (left columns) and their full names (right columns) are shown. Note Rows in red correspond to biomass precursors in the central carbon metabolism. atp, nadph and nad are also biomass precursors, but we wish to emphasize on metabolites that are act as biochemical precursors to the actual biomass precursors of *E. coli* (See main text). atp, nadph and nad are also biomass precursors for *E. coli*.

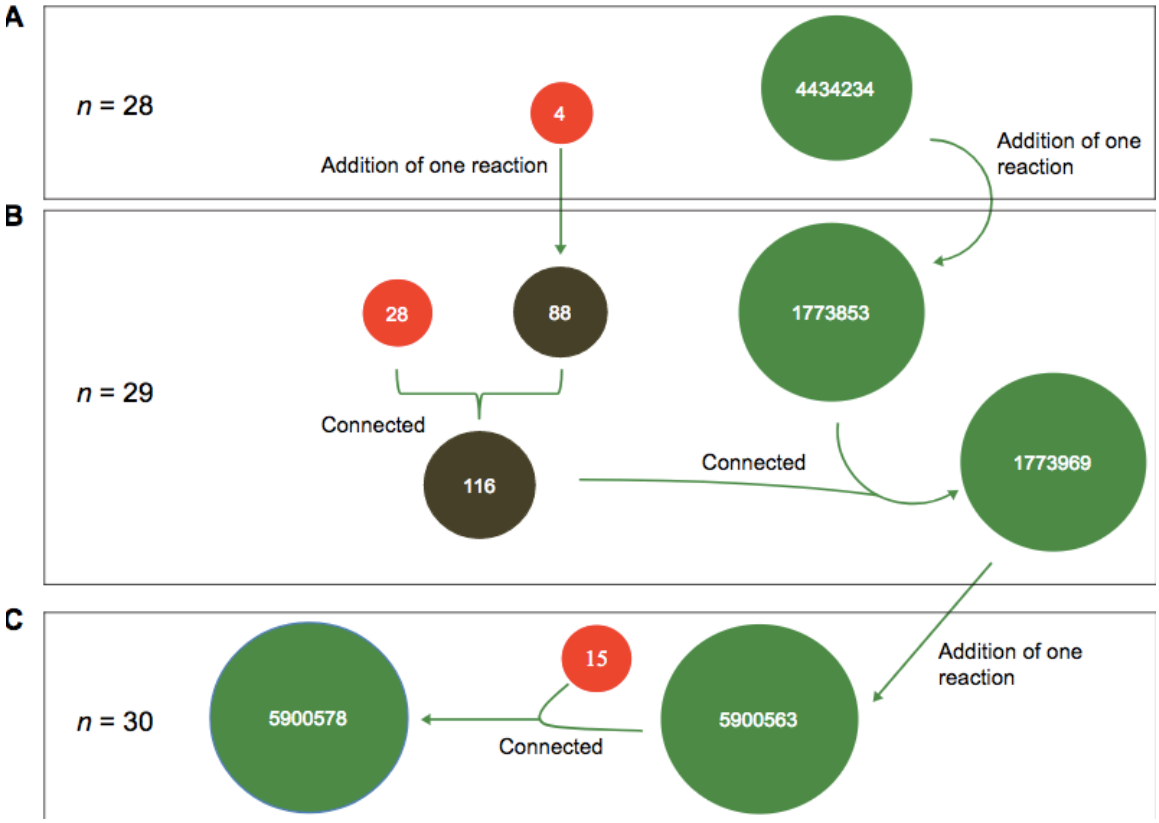
| Metabolism size $n$ | Number of metabolisms $\Omega(n)$ | Number of viable potential metabolisms $V(n)$ on glucose | Fraction of viable potential metabolisms on gluose |
|---|---|---|---|
| 23 | 1.9679E+14 | 3 | 1.52444E-14 |
| 24 | 2.2959E+14 | 91 | 3.96355E-13 |
| 25 | 2.4796E+14 | 1333 | 5.37588E-12 |
| 26 | 2.4796E+14 | 1.2512E+04 | 5.04599E-11 |
| 27 | 2.2959E+14 | 8.4344E+04 | 3.67365E-10 |
| 28 | 1.9679E+14 | 4.3424E+05 | 2.20657E-09 |
| 29 | 1.5608E+14 | 1.7740E+06 | 1.1366E-08 |
| 30 | 1.1446E+14 | 5.9006E+06 | 5.15529E-08 |
| 31 | 7.7535E+13 | 1.6274E+07 | 2.09889E-07 |
| 32 | 4.8459E+13 | 3.7712E+07 | 7.78212E-07 |
| 33 | 2.7901E+13 | 7.4145E+07 | 2.65744E-06 |
| 34 | 1.4771E+13 | 1.2456E+08 | 8.43246E-06 |
| 35 | 7.1745E+12 | 1.7967E+08 | 2.50429E-05 |
| 36 | 3.1887E+12 | 2.2325E+08 | 7.0013E-05 |
| 37 | 1.2927E+12 | 2.3933E+08 | 0.000185138 |
| 38 | 4.7626E+11 | 2.2138E+08 | 0.00046484 |
| 39 | 1.5875E+11 | 1.7647E+08 | 0.0011116 |
| 40 | 4.7626E+10 | 1.2087E+08 | 0.002537936 |
| 41 | 1.2778E+10 | 7.0817E+07 | 0.005542198 |
| 42 | 3.0423E+09 | 3.5259E+07 | 0.01158958 |
| 43 | 6.3676E+08 | 1.4786E+07 | 0.023220727 |
| 44 | 1.1578E+08 | 5.1600E+06 | 0.044569549 |
| 45 | 1.8009E+07 | 1.4745E+06 | 0.081871194 |
| 46 | 2.3491E+06 | 3.3744E+05 | 0.143647672 |
| 47 | 2.4990E+05 | 5.9966E+04 | 0.239959984 |
| 48 | 2.0825E+04 | 7909 | 0.379783914 |
| 49 | 1275 | 721 | 0.565490196 |
| 50 | 51 | 40 | 0.784313725 |
| 51 | 1 | 1 | 1 |

**Additional File 3: Number of potential metabolisms viable on all ten carbon sources as a function of metabolism size.**



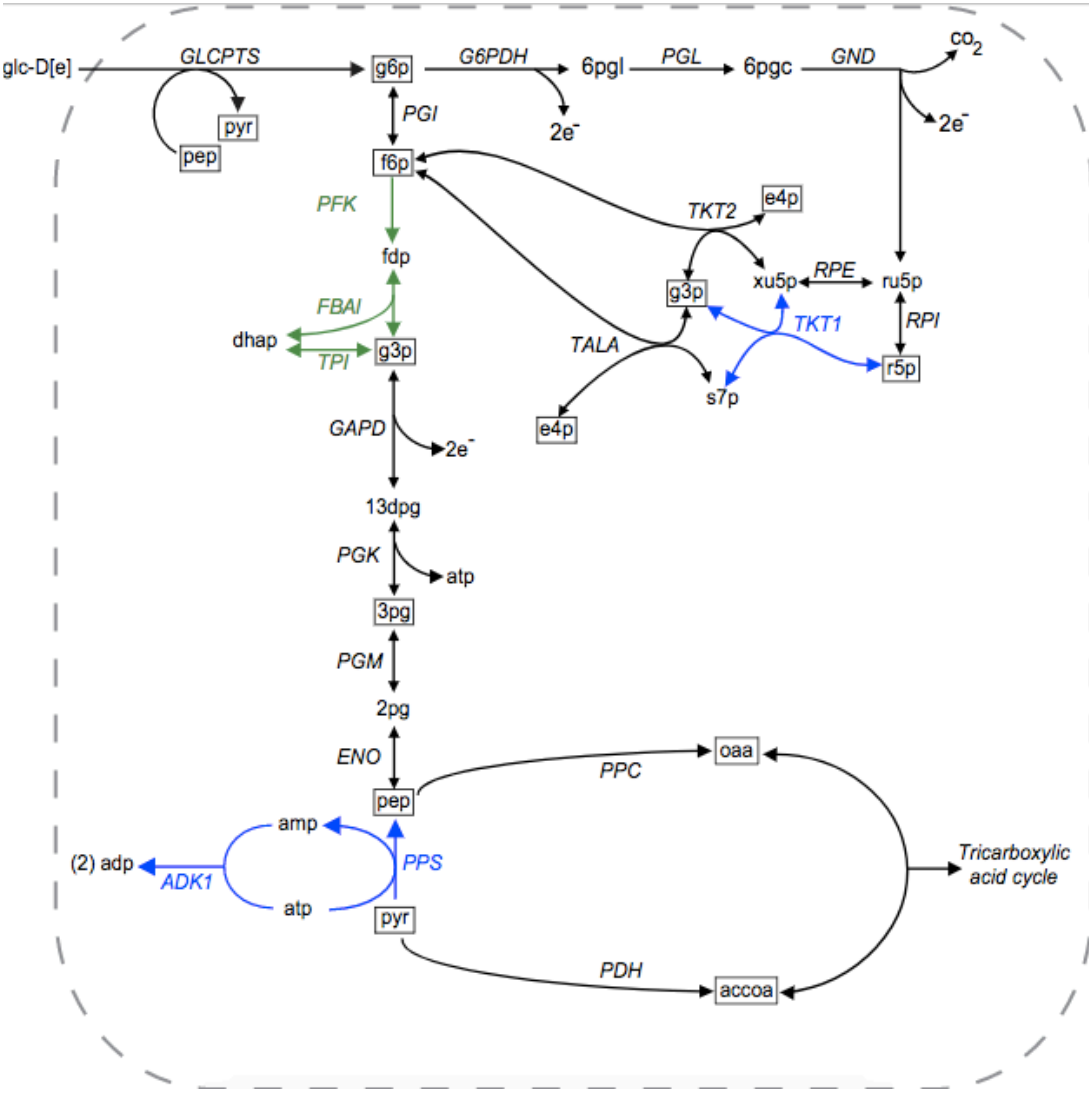

**Additional file 4: An example of a minimal metabolism viable on glucose.** The figure shows an example of a minimal metabolism of size 23, which is also one of the smallest metabolisms viable on glucose. All 13 biomass precursors are framed with solid rectangles. Only important transport reactions and cofactors are shown. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Abbreviations are spelled out in Additional file 2.

**Additional file 5: Supplementary results.** Downloadable at https://static-content.springer.com/esm/art%3A10.1186%2F1752-0509-8-48/MediaObjects/12918_2014_1319_MOESM5_ESM.pdf




**Additional file 6: Connectedness of genotype networks containing viable genotypes of** n **= 28, 29, and 30 reactions.** The figure shows the connectivity of genotype networks for reactions in central carbon metabolisms, as a function of size *n*. Each circle corresponds to a connected component, and the number in each circle corresponds to the number of genotypes in this component. The components at size 28 were obtained by full enumeration, but for larger sizes such an approach is not feasible. Instead one has to use a form of recursive evaluation that we illustrate here for two larger sizes. Panel **(A)** shows the two disconnected components in the network corresponding to size 28, one containing 434234 genotypes, and the other component containing just 4 genotypes. The addition of one reaction to these 4 genotypes results in 88 genotypes of size 29, which must be connected (see main text). **(B)** Aside from these 88 connected genotypes, there are also 28 minimal genotypes of size 29 (red circles). We verified computationally that both groups of genotypes (88 and 29) were connected using breadth-first search and found that they form a single component of 116 genotypes. We were also able to demonstrate that this component is connected to the 1773853 connected genotypes that are parents of the large component at size $n = 28$ (panel A). The two thus form a connected

genotype network of 1773969 metabolisms of size 29. Panel **(C)** shows that adding one reaction to these genotypes results in 5900563 connected genotypes at size 30. In addition, 15 new minimal metabolisms (red circles) come into being at size 30. We found that they were connected to the remaining 5900563 genotypes, thus forming a single connected network comprising 5900578 genotypes. (PDF 12 KB)




**Additional file 7: An example of essential pathways that result in fragmentation of genotype space.** The figure shows the two essential pathways (green and blue) in a pair of genotypes belonging to the two disconnected components in the genotype network of potential metabolisms with $n = 25$ reactions (Figure 3C). The reactions in green, catalyzed by enzymes PFK, FBAl and TPI are essential in all genotypes belonging to the largest component (subgraphs *A"* and *B"*) in Figure 3C, while the reactions in blue are essential to all four genotypes in component *C* in Figure 3C. The

13 biomass precursors are surrounded by black rectangles. Only important transport reactions and cofactors have been shown. Enzymes catalyzing each of the reactions are shown in uppercase italic typeface. Information on abbreviations is provided in Additional file 2.

| **Metabolism size *n*** | **Number of metabolisms Ω(*n*)** | **Number of potential metabolisms viable on all carbon sources** | **Fraction of potential metabolisms viable on all carbon sources** |
|---|---|---|---|
| 34 | 1.4771E+13 | 4 | 2.708E-13 |
| 35 | 7.1745E+12 | 79 | 1.10112E-11 |
| 36 | 3.1887E+12 | 736 | 2.30817E-10 |
| 37 | 1.2927E+12 | 4249 | 3.2869E-09 |
| 38 | 4.7626E+11 | 1.6869E+04 | 3.54197E-08 |
| 39 | 1.5875E+11 | 4.8507E+04 | 3.05549E-07 |
| 40 | 4.7626E+10 | 1.0399E+05 | 2.18349E-06 |
| 41 | 1.2778E+10 | 1.6902E+05 | 1.32277E-05 |
| 42 | 3.0423E+09 | 2.1017E+05 | 6.90827E-05 |
| 43 | 6.3676E+08 | 2.0060E+05 | 0.000315025 |
| 44 | 1.1578E+08 | 1.4667E+05 | 0.001266835 |
| 45 | 1.8009E+07 | 8.1539E+04 | 0.004527565 |
| 46 | 2.3491E+06 | 3.3985E+04 | 0.014467489 |
| 47 | 2.4990E+05 | 1.0377E+04 | 0.04152461 |
| 48 | 2.0825E+04 | 2237 | 0.107418968 |
| 49 | 1275 | 320 | 0.250980392 |
| 50 | 51 | 27 | 0.529411765 |
| 51 | 1 | 1 | 1 |

**Additional file 8: Number of metabolisms viable on all ten carbon sources as a function of metabolism size.**

**Additional file 9: The essential computer programs used in this analysis.**
Downloadable at https://static-content.springer.com/esm/art%3A10.1186%2F1752-0509-8-48/MediaObjects/12918_2014_1319_MOESM9_ESM.zip

# Chapter 6:

# Exhaustive analysis of a genotype space comprising $10^{15}$ central carbon metabolisms reveals an organization conducive to metabolic innovation

Sayed-Rzgar Hosseini, Aditya Barve and Andreas Wagner

The content of this chapter has been published as:

Hosseini, S.-R., A. Barve, and A. Wagner. 2015. Exhaustive analysis of a genotype space comprising $10^{15}$ central carbon metabolisms reveals an organization conducive to metabolic innovation. PLoS Comput. Biol. 11: e1004329. https://doi.org/10.1371/journal.pcbi.1004329

## 6.1. Abstract

All biological evolution takes place in a space of possible genotypes and their phenotypes. The structure of this space defines the evolutionary potential and limitations of an evolving system. Metabolism is one of the most ancient and fundamental evolving systems, sustaining life by extracting energy from extracellular nutrients. Here we study metabolism's potential for innovation by analyzing an exhaustive genotype-phenotype map for a space of $10^{15}$ metabolisms that encodes all possible subsets of 51 reactions in central carbon metabolism. Using flux balance analysis, we predict the viability of these metabolisms on 10 different carbon sources which give rise to 1024 potential metabolic phenotypes. Although viable metabolisms with any one phenotype comprise a tiny fraction of genotype space, their absolute numbers exceed $10^9$ for some phenotypes. Metabolisms with any one phenotype typically form a single network of genotypes that extends far or all the way through metabolic genotype space, where any two genotypes can be reached from each other through a series of single reaction changes. The minimal distance of genotype networks associated with different phenotypes is small, such that one can reach metabolisms with novel phenotypes – viable on new carbon sources – through one or few genotypic changes. Exceptions to these principles exist for those metabolisms whose complexity (number of reactions) is close to the minimum needed for viability. Increasing metabolic complexity enhances the potential for both evolutionary conservation and evolutionary innovation.

## Author Summary:

Genotype-phenotype mapping provides an unprecedented opportunity to gain new insights into the function of biological systems and their evolution. We present a comprehensive genotype-phenotype map for a genotype space comprising more than $10^{15}$ central carbon metabolisms. The subsets of viable metabolisms form a connected genotype network that extends far through genotype space, and that renders multiple novel phenotypes in the immediate neighborhood of viable metabolisms accessible. The map we describe reveals an organization of core metabolism that simultaneously facilitates evolutionary conservation of existing metabolic phenotypes and the origination of novel metabolic traits that allow viability on novel carbon sources.

## 6.2. Introduction

Attempts to understand the relationship between genotype and phenotype have played a pivotal role in the history of genetics and evolutionary biology, beginning with the rediscoveries of Mendel's laws, which revealed genotypes as distinct from phenotypes and responsible for the inheritance of phenotypic traits [1]. Subsequent efforts to map individual genes encoding such traits onto chromosomes helped develop classical and molecular genetic mapping methods [2]. Decades later, genome sequencing facilitated more systematic studies of the relationship between genotypes and complex traits, and resulted in the emergence of functional genomics and systems biology. Today, a central goal of systems biology is to uncover and predict the relationship between genotypes and complex phenotypes, which include the structure and dynamics of complex intracellular networks like gene regulatory, signaling, and metabolic networks [3,4].

Computational methods to predict phenotype from genotype are increasingly powerful and they can help establish genotype-phenotype maps at unprecedented resolution [5–7]. An ideal such map would cover all possible genotypes, but unfortunately, the entire space of a system's genotypes is usually too vast to study. For instance, the space of all RNA sequences of length 100 comprises $4^{100}= 10^{60}$ different RNA sequences, and the space of all protein sequences of length 100 comprises $20^{100}=10^{130}$ different protein sequences. Exhaustive genotype-phenotype mapping using computational approaches is possible only in small systems. They include short hydrophobic polar (HP) model proteins folding on square and cubic lattices [8], short RNA sequences folding into planar secondary structures [9], and small gene regulatory networks [10]. Similar exhaustive approaches have not been taken yet in metabolic systems, and our understanding of metabolic systems is thus far limited to sampling of metabolic genotype space [11–18]. Here, we build on our previous work studying the connectivity of the space of central carbon metabolisms [19] and present an exhaustive genotype-phenotype map of this space.

In evolution, phenotypic change is caused by genotypic change. Thus, the structure of a genotype-phenotype map contains information about the evolutionary potential and limitations of an evolving system. Metabolisms are important classes of evolving systems, because they are a source of many evolutionary adaptions and innovations, especially in microorganisms. For instance, microorganisms have acquired the ability

to utilize many non-natural substances like polychlorinated biphenyls, chlorobenzens, organic solvents and synthetic pesticides as food [20–23]. Microbial isolates from pristine soils can survive on several antibiotics like ciprofloxacin by using them as sole carbon sources [24], and halophilic bacteria can tolerate high salt concentration by synthesizing novel molecules like ectoine or glycine betaine [25,26].

An organism's metabolism is a biochemical reaction network that comprises a set of reactions that are catalyzed by enzymes encoded by genes. Following common practice in the computational analysis of metabolism [4], we here represent a metabolic genotype as a binary presence-absence pattern of a set of biochemical reactions, i.e., as a binary string whose entries indicate presence (1) or absence (0) of a reaction among some set of $N$ possible reactions. In our analysis (see also [19]), this global reaction set comprises *51* reactions from central carbon metabolism (see Methods, and table S9). Any one metabolism can be viewed as a point in metabolic genotype space whose size is equal to $2^{51}=2.25 \times 10^{15}$ metabolisms. In this framework, a metabolism's genotype can evolve by either losing one or more reactions, for example through a loss of function mutation, or by gaining one or more reactions, for example through horizontal gene transfer or gain of function mutations [27–29].

We focus on central carbon metabolism because of its pivotal role in extracting energy from extracellular carbon sources. It comprises the interconnected biochemical pathways of glycolysis, gluconeogenesis, the pentose-phosphate pathway (PP), and the tricarboxylic acid cycle (TCA). These are supplemented by anaplerotic reactions and the glyoxylate shunt [30]. Glycolysis converts glucose into pyruvate and produces high-energy compounds like ATP and NADH. In parallel, the pentose-phosphate pathway generates NADPH and pentose sugars required for anabolic reactions. The glycolytic end product pyruvate is oxidized to acetyl-CoA, which enters the tri-carboxylic acid cycle (TCA), a cyclical series of reactions that generate ATP, NADH, and amino acid precursors. Pyruvate and phosphoenolpyruvate (PEP) can also enter the TCA cycle directly through anaplerotic reactions that replenish TCA cycle intermediates consumed in biosynthetic processes. Conversely, under gluconeogenic conditions, the TCA cycle intermediates oxaloacetate or malate are converted to pyruvate and PEP to provide the precursors for gluconeogenesis. Acetyl-CoA can also participate in the glyoxylate shunt to

generate succinate for carbohydrate synthesis. Finally, reactions of the oxidative phosphorylation pathway participate in production of ATP from NADH (Figure S1).

Not all organisms contain all reactions associated with central carbon metabolism. For example, although the Embden–Meyerhoff–Parnass (EMP) glycolytic pathway (figure S1) is nearly ubiquitous among eukaryotes, prokaryotes can use diverse natural glycolytic alternatives [31–33]. Archaebacteria like Sulfolobus solfataricus [34] and Thermoplasma acidophilum [35], metabolize glucose through the Entner-Doudoroff (ED) pathway and heterolactic fermentative bacteria use phosphoketolase pathway instead of the canonical glycolytic pathway [32]. Natural glycolytic pathways vary in their reaction content and in how much ATP they produce per glucose [33,36]. Furthermore, genome analysis of 19 species including 4 archaea, 14 bacteria and 1 eukaryote has revealed that in the majority of species the citric acid cycle is incomplete or absent [37]. Also, in vivo quantification of intracellular carbon fluxes from $C_{13}$ tracer experiments has shown that relative activities of individual reactions of central carbon metabolism vary widely among different species [38]. Such variation calls for an examination of how genotypic variation maps into variation in metabolic phenotypes.

For our analysis, we define a metabolic phenotype based on a metabolism's *viability*, its ability to sustain life in a given set of environments. We call a metabolism viable in any one environment if it can synthesize 13 key precursors (Figure S1) that are produced by central carbon metabolism and that are necessary for the synthesis of amino acids and other essential biomass molecules [39]. We here consider 10 environments that vary in their carbon source and contain an otherwise minimal complement of nutrients (see Methods). We represent the phenotype of a given metabolism as a binary string of length 10 whose entries indicate viability (1) or inviability (0) of the metabolism on a given carbon source. In this framework, a change in metabolic genotype is a metabolic innovation if it leads to viability on new combinations of carbon sources. To compute the phenotype of a given metabolic genotype, we use the constraint-based method of flux balance analysis (FBA, see Methods) [7,40–42] whose qualitative predictions of viability are in good agreement with experimental data [43–49]. FBA takes advantage of constraints imposed by the stoichiometry of all metabolic reactions and maximal nutrient uptake rates in a given environment. Subject to the law of mass conservation in a metabolic steady state, it

can predict the rate at which each reaction can proceed under conditions of maximal production of biomass precursors [7,40,44,47,50–52]. Although FBA is computationally efficient [7,40–42], determining the phenotypes of more than $10^{15}$ metabolic genotypes in each of 10 possible environments is challenging and required us to develop techniques that simplify this task (see Methods).

We use the resulting genotype-phenotype map to analyze a set of viable metabolisms with a given phenotype as a graph, where two nodes (metabolisms) are neighbors if they can be connected by a single reaction change. We study the organization of these graphs in metabolic genotype space, and find that they typically extend far through this space. To study a metabolism's potential for metabolic innovation, that is, how changes in metabolic reactions can lead to new metabolic phenotypes (viability in new environments) we explore the metabolism's neighborhood in genotype space and this neighborhood's phenotypic diversity. We find that the neighborhoods of metabolisms with the same phenotype but different genotype often contain metabolisms with different novel metabolic phenotypes. Also, the genotype networks of different metabolic phenotypes abut each other in genotype space. Together, these observations suggest that the organization of metabolic genotype space is conducive to metabolic innovation.

## 6.3. Results

### The number of viable metabolisms depends on carbon source and reaction numbers

We first enumerated the number of central carbon metabolisms (CCM) on a given carbon source. Figure 1a shows this number as a function of metabolism size, that is, the number $n$ of reactions in a metabolism, for various carbon sources. Note the logarithmic vertical axis. Black data points indicate the total number of metabolisms of a given size $n$, regardless of their viability. This number is given by the binomial coefficient $\binom{N}{n}$. The following observations emerge from this figure. First, the minimum number of reactions $n_{\min}$ in a viable metabolism is not the same for all carbon sources. For example, it varies from $n_{min}=23$ for glucose and fructose to $n_{min}=30$ for acetate (Figure S2, See the Supplementary Text S1 for an explanation of this difference).

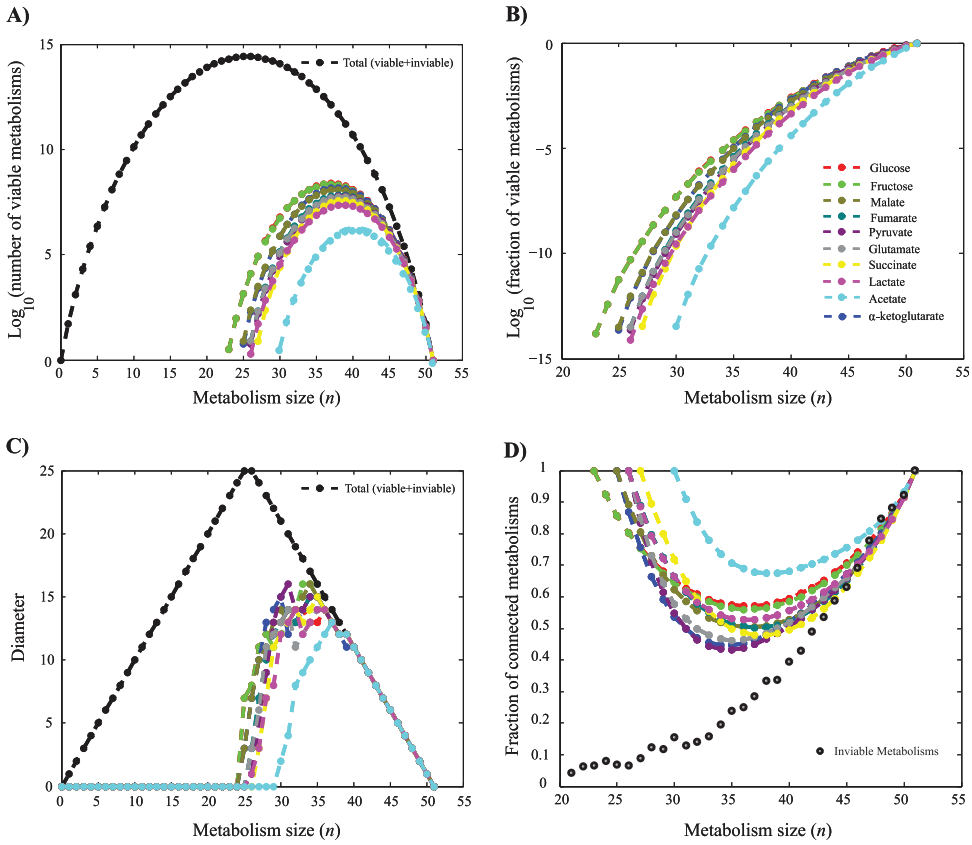



**Figure 1: Metabolisms viable on different carbon sources and genotypic differences among them.** **a)** Number of viable metabolisms. Black circles (vertical axis) indicate the total possible numbers $\binom{N}{n}$ of metabolisms of a given size $n$ (horizontal axis, $N$=51, $0 \leq n \leq 51$). Colored data points indicate the number of metabolisms viable on the single carbon sources indicated in the legend. **b)** Fraction of metabolisms viable on different carbon sources. Note the logarithmic vertical scale. Data on glucose in a) and b) has been previously published [19]. **c)** Genotype network diameter**.** Black circles indicate the diameter of genotype space for metabolisms of a given size, which is an upper bound to the diameter of any one genotype network. Colored data points indicate genotype network diameter for metabolisms viable on a single carbon source. Where sets of viable metabolisms comprised more than one connected component, the diameter of the giant component was chosen. At $n \geq 40$, all genotype networks have the maximally possible diameter. **d)**. Fraction of metabolisms that contain no disconnected reactions. Colored circles correspond to metabolisms viable on the carbon source specified in the legend and black circles correspond to inviable metabolisms. Data on inviable metabolisms is based on 10000 randomly sampled inviable metabolisms. Interpolation between data points (circles) is linear and displayed as a visual guide.

Second, the maximum number of viable metabolisms also varies by more than two orders of magnitude, from approximately $1.9\times10^6$ for acetate to $2.4\times10^8$ for glucose (see Table S1).

Third, the number of viable metabolisms shows a unimodal distribution whose qualitative shape is predicted by binomial coefficients that are shifted by $n_{\min}$, i.e., $\binom{N-n_{\min}}{n-n_{\min}}$, for any one carbon source (See figure S3a and S3b for examples). This is a result of the fact that adding any number of reactions to a viable metabolism of minimal size $n_{\min}$ creates another viable metabolism, and there are $\binom{N-n_{\min}}{n-n_{\min}}$ ways of adding $n - n_{\min}$ reactions. We note that this qualitative relationship is not quantitatively accurate, for reasons explained in the supplementary Text S2.

Figure 1b shows the same information as figure 1a, but expresses the number of viable metabolisms as a fraction of all metabolisms $\binom{N}{n}$ at a given size *N*. Note the logarithmic scale and that the fraction of viable metabolisms declines faster than exponentially with decreasing *n*. It is easy to understand this pattern if one considers a metabolism ("the child") that is derived from another ("the parent") by eliminating a single reaction. First, some fraction of the children of viable parent metabolisms is inviable; second, this fraction increases as *n* decreases; third, all children of inviable metabolisms are themselves inviable. Together, these observations can account for the rapid decrease of the fraction of viable metabolisms.

**Except for the smallest metabolisms, most viable metabolisms form a genotype network with a large diameter.**

Viable genotypes of all but the smallest sizes *n* form a single connected network of genotypes, where any two genotypes can be reached from one another through changes in individual reactions. In a previous contribution, we have demonstrated this connectivity for metabolisms viable on glucose and on all ten carbon sources we consider here taken together [19]. Table S2 shows that it also holds for metabolisms viable on the nine other carbon sources. Even where the set of viable metabolisms is partitioned into more than one component ($n_C>1$; table S2), i.e., more than one genotype network, this number of components is usually small, and the largest of them harbors a large fraction $r_G$ of genotypes ($r_G>0.95$ in 23 out of 29 cases where $r_G<1$; table S2).

Starting from these observations, we asked how different the reaction content of metabolisms viable on the same carbon source could be. In graph theoretic terms, this is a question about the diameter of the set of viable genotypes at any one size *n*. Figure 1c shows this diameter as a function of metabolism size *n* for all 10 carbon sources. For most metabolism sizes, the diameter lies between 5 and 15, indicating that metabolisms of the same size and viable on the same carbon sources can have very different reaction complements. Moreover, for any one carbon source, and for metabolisms of most sizes above $n_{min}$, the diameter as a fraction of the diameter of the genotype space (i.e. the maximum possible diameter (See methods)) is equal to one or very close to one (figure 1c). This implies that the set of viable metabolisms is not highly localized within a small region of genotype space. Rather, its members occur throughout the space and the set comprising them often spans this space. Since the fraction of viable metabolisms increases with *n*, the fractional diameter increases with increasing *n* (until *n*=40, where it reaches one for all carbon sources).

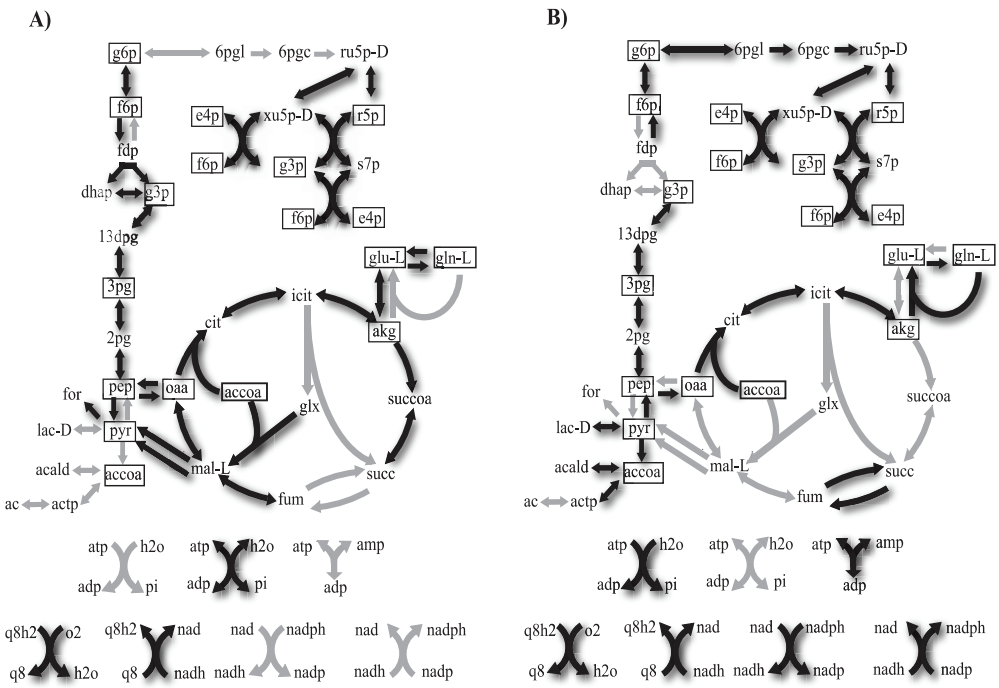



**Figure 2: Example of two maximally different metabolisms.** Each arrow in each panel corresponds to one of the 51 internal reactions we consider. Black arrows and gray arrows correspond to reactions that are present or absent, respectively, in the metabolism shown. Metabolites are indicated by their acronyms (see Table S9). Boxed metabolites correspond to 13 essential biomass precursors. Note that 4 metabolites (accoa, g3p, f6p and e4p) are shown more than once for visual clarity. **a)** A metabolism with 33 reactions that is viable on fructose and that differs from another metabolism of size 33 viable on fructose shown in **b)** by 16 reactions. The two metabolisms share 17 reactions.

As an example of two viable networks comprising maximally different reaction sets, figure 2 shows two maximally different metabolisms viable on fructose. Each metabolism contains 33 reactions. The two metabolisms share 17 reactions, and differ in 16 reactions. The products of the 17 shared reactions are the essential biomass precursors (boxed molecules in figure 2), meaning that the respective reactions cannot be bypassed via alternative reactions or pathways. In contrast, the 16 differing reactions are either non-essential reactions that do not directly contribute to the production of small biomass molecules, or they are reactions essential only in one of the metabolisms but not in both. Such non-shared but essential reactions exist, because a necessary biomass molecule may be synthesized by one or more alternative reactions or pathway in two metabolism. An example is the reaction catalyzed by glutaminase, which is essential for the synthesis of glutamate (glu-L) from glutamine (gln-L) and $\alpha$-ketoglutarate (akg) in the metabolism of figure 2b, but is substituted for by the reaction catalyzed by glutamate dehydrogenase, which is essential for production of glutamate from $\alpha$-ketoglutarate (akg) and ammonium (not shown) in the metabolism of figure 2a.

Central carbon metabolism contains pairs of reactions that differ only in the co-factor they use. Three such pairs are relevant for our analysis. The first comprises two reactions catalyzed by malic enzyme that use NAD and NADP as co-factors. The second comprises two reactions catalyzed by different transhydrogenases, and the third includes the reactions catalyzed by ATPase and ATP synthase (Table S9). We wanted to examine whether such pairs of reactions trivially increase the genotypic distance among metabolism pairs and thus artificially inflate genotype network diameter. For any given genotype network, this could only be the case if the two members of all metabolism pairs with the highest genotypic distance use different reactions from at least one of these two pairs. To find out, we first enumerated all metabolism pairs with genotypic distance equal to the diameter of the given genotype network, and noted that there are up to millions of such pairs. Next, we examined for each such pair of metabolisms whether (i) one member used the NAD isoform of malic enzyme while the other one used the NADP isoform, or (ii) one member used one transhydrogenase while the other member used the other transhydrogenase or (iii) one member used ATPase and the other one ATP synthase. If so, we eliminated the pair from the set of pairs with largest distance. Table S3 shows the percentage of

metabolism pairs that remained in the set based on this criterion. It exceeds 30% for most genotype networks and is greater than zero for every genotype network. This observation implies that in every genotype network at least one metabolism pair does not have artificially large genotypic distance due to co-factor dependency. Hence, co-factor dependency does not artificially inflate genotype network diameter.

Reactions that are blocked – they must have zero flux for stoichiometric reasons [11, 53] – could also contribute to large genotype network diameter. To find out whether this is the case, we calculated the total number of blocked reactions for each metabolism pair that is viable on the same carbon source and whose genotypic distance is equal to the diameter of its genotype network. Table S4 shows, for each carbon source, the minimum number of blocked reactions among such metabolism pairs. For most genotype networks (71.6% percent, 189 of 253) this minimum is zero. This means that at least one pair of metabolisms without any blocked reaction exists among the metabolism pairs with the largest genotypic distance. The exceptions are some genotype networks of metabolisms with intermediate sizes, where some contain no metabolism pairs without blocked reactions. Even in these genotype networks, however, only one or at most two reactions are blocked, meaning that the genotype network diameter would decrease only by this number if one were to disregard blocked reactions.

**The majority of reactions are connected to one another in most viable metabolisms.**

In a horizontal gene transfer event, enzyme-coding genes can be imported into a genome whose products catalyze reactions that may be connected to or disconnected from the resident metabolism. We define a reaction as disconnected from a metabolism if (i) its products are neither biomass precursors nor substrates of any other reaction in the resident metabolism, or (ii) at least one of its substrates is neither a product of other reactions nor a nutrient taken up from the environment. For example, among the 16 reactions that differ between the metabolisms in figure 2, one reaction in each of the two metabolisms is disconnected from the rest, that is, its substrates are neither products of other reactions nor are they nutrients provided by the environment in which fructose is the sole carbon source. Specifically, in figure 2a, the reaction that is catalyzed by malate synthase and produces L-malate (mal-L) from

glyoxylate (glx) and acetyl-coenzyme A (accoa) is a disconnected reaction, because glyoxylate is neither available in the environment, nor is it produced by other reactions in the metabolism. Similarly, in figure 2b the reaction that is catalyzed by fructose-bisphosphatase and produces fructose 6-phosphate (f6p) from fructose 1,6-bisphosphate (fdp) is disconnected.

To find out whether such disconnected metabolisms could strongly influence our analysis of metabolic genotype space, we determined how abundant they are. Specifically, we first computed the fraction $f_d$ of all viable metabolisms that contained at least one such disconnected reaction. The value of $f_d$ ranged from 0.517 for metabolisms viable on pyruvate to 0.307 for those viable on acetate. Figure 1d shows 1- $f_d$, i.e., the fraction of viable metabolisms containing only connected reactions as a function of metabolism size. We note that for any one carbon source, the smallest viable metabolisms contain only connected reactions. It is easy to see why this must be the case: if a minimal metabolism contained a disconnected reaction, then this reaction would by definition be dispensable, and its elimination could not abolish viability, which means that the metabolism could not possibly be minimal. Conversely, the largest viable metabolism contains all reactions we consider, which are also connected. Only at intermediate sizes do metabolisms with disconnected reactions occur. However, we also found that at most sizes, most viable metabolisms contain only connected reactions, regardless of the carbon source considered (Figure 1d). This stands in contrast to the fraction of *inviable* metabolisms lacking disconnected reactions (i.e. $(1-f_d)_n$ (black circles in Figure 1d), which is much smaller for metabolisms up to about 42 reactions, where it approaches that of viable metabolisms.

Given these observation, it is thus of little surprise that the patterns we reported above extend to metabolisms where all reactions are connected. Specifically, the quasi-binomial dependence of the number of viable genotypes on $n$, and the greater than exponential reduction in the fraction of viable genotypes with decreasing $n$, are preserved (Figures S4a and S4b). Moreover, such metabolisms can also be quite different from one another (Figure S4c) and at most sizes $n$, most or all such metabolisms reside in a single connected genotype network (Table S5).

**Metabolisms viable on multiple carbon sources show similar organization**

In a next analysis, we asked how the observations we made so far translate into metabolisms that are viable on some number $k>1$ of carbon sources. This analysis is more challenging, because at each $k$ there are $\binom{10}{k}$ possible $k$-tuples of carbon sources. We exhaustively enumerated, for each possible $k$-tuple of carbon sources, the number of metabolisms of a given size that are viable on that $k$-tuple. Then, we calculated the average number of metabolisms viable on a given $k$-tuple. Figure 3a and 3b show the number and fraction of viable metabolisms as a function of metabolism size $n$, and for different values of $k$. Several observations are germane. First, the unimodal relationship between the number of viable genotypes and metabolism size still holds for metabolisms viable on multiple carbon sources. Second, the smallest number of reactions in a viable metabolism increases from 23 for metabolisms viable on a single carbon source to 34 for metabolisms viable on 10 carbon sources. Third, the fraction of viable metabolisms declines at a greater than exponential rate as the number of reactions decreases (Figure 3b). Fourth, the rate of this decline becomes steeper as the number of carbon sources increases on which a metabolism is viable (Figure 3b).

Tables S6, S7, and S8, respectively, show the median, maximum, and minimum of the number $n_C$ of connected components, and the fraction of metabolisms in the largest ("giant") component for metabolisms of size $n$ required to be viable on $k$ carbon sources. The results show that for metabolisms above $n$=36 reactions, all metabolisms viable on a given number of carbon sources are connected. Wherever the set of viable metabolisms are disconnected, they are partitioned into few connected components (with a median of 2-3 and a maximum of 7), and with few exceptions most metabolisms reside in the largest of these components.

Figure 3c shows the median diameter of the set of genotypes viable on $k$ carbon sources as a function of $n$ (see figure S5a, and S5b, for the minimum and maximum diameters for each $k$). This median diameter is not substantially lower than for metabolisms viable on single carbon sources. Specifically, for metabolisms of most sizes, the diameter of the set of viable metabolisms lies between 5 and 15, indicating that metabolisms of the same size and viable on the same number of carbon sources can differ substantially in their reaction complement. Moreover, the median diameter

as a fraction of the maximally possible diameter, i.e., the diameter of genotype space, is one or close to one for most sizes above $n_{min}$. In other words, the set of viable metabolisms viable on any set of carbon sources is not localized to a small region of genotype space. It often spans the entire space.

Figure 3d shows not only that the majority of metabolisms considered lack disconnected reactions, but also that the fraction of metabolisms without any disconnected reactions increases with the number $k$ of carbon sources on which viability is required. The patterns of organization from figure 3a to 3c also hold for viable metabolisms where all reactions are biochemically connected to one another, as shown in figures S6 and S7.

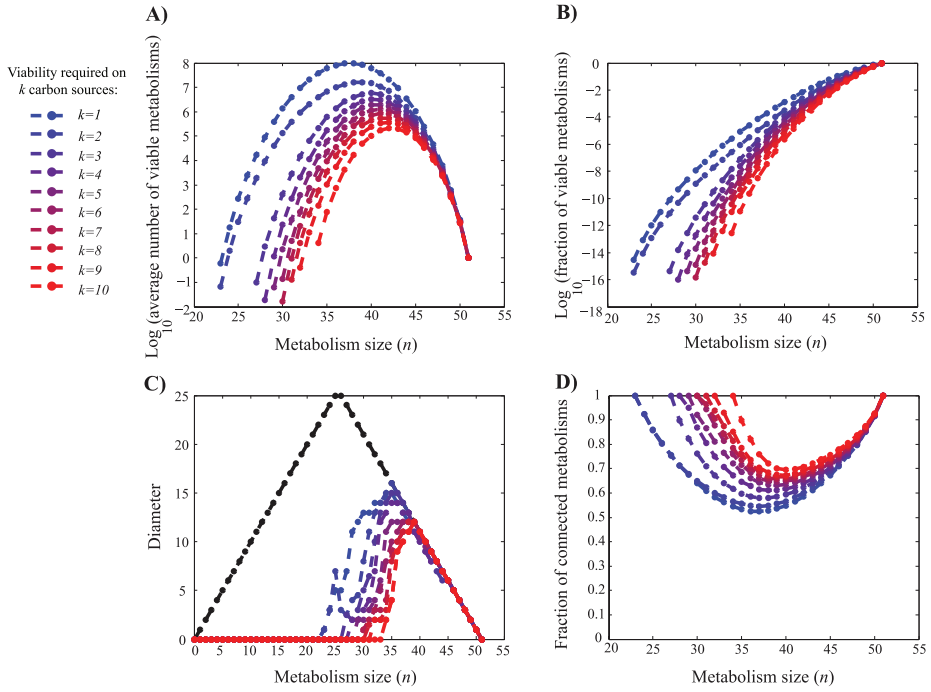



**Figure 3: Metabolisms viable on multiple carbon sources and genotypic differences among them.** **a)** Average number of metabolisms per genotype network that are viable on $k$ carbon sources for $1 \leq k \leq 10$. Each curve corresponds to one value of $k$ and is colored as indicated in the legend. **b)** Fraction of metabolisms viable on $k$ carbon sources. Note the logarithmic scale. Data on ten carbon sources in a) and b) has been previously published [19]. **c)** Genotype network diameter**.** Black circles indicate the diameter of genotype space for metabolisms of a given size, which is an upper bound to the diameter of any one genotype network. Colored data points indicate the median of the diameter of the genotype networks of metabolisms viable on $k$ carbon sources. At $n \geq 40$, almost all genotype networks have the maximally possible diameter. **d)** Fraction of metabolisms that are viable on $k$ carbon sources and contain no disconnected reactions.

**Local neighborhoods of viable metabolisms are phenotypically diverse**

We refer to the neighborhood of a metabolism *M* as the set of metabolisms differing in one reaction from *M*. Neighborhoods are important in the evolution of biological systems, because they contain those genotypes that are easily reachable through a small genotypic change – in our case, change of a single reaction- from a given genotype. In our next analysis, we studied the number of novel phenotypes contained in such neighborhoods. To this end, we first sampled 1000 metabolisms of a given size from each of the distinct genotype networks viable on different carbon sources. Representing each phenotype as a binary vector of length 10 whose *i*-th entry indicates viability (1) or inviability (0) on the *i*-th carbon source, we asked whether multiple distinct novel phenotypes in the neighborhood of a given metabolism *M* exist, i.e. phenotypes different from *M* that indicate growth on at least one additional carbon source compared to that of *M*. The answer is yes. (Figure 4a). Supplementary Text S3 explains the single-peaked shape of the distributions in Figure 4a. Figure S8 shows that the number of these novel accessible phenotypes is greater than expected by chance, based on a simple randomization test.

In a next analysis, we asked whether the neighborhoods of different metabolisms viable on the same carbon source contain different novel phenotypes. If so, which novel phenotypes are accessible may depend on where in metabolic genotype space a viable metabolism is located. To find out, we randomly and uniformly sampled 1000 pairs of metabolisms with *n* reactions and viable on a given carbon source, where each pair differed in *D* reactions. Then, we determined the set of new phenotypes accessible in the local neighborhood of metabolisms $M_1$ and $M_2$, which we denote by $P_1$ and $P_2$, respectively. We then computed the fraction *u* (for *unique*) phenotypes, which appear in $P_1$ or $P_2$ but not in both, i.e., $u=1-|P_1 \cap P_2|/(|P_1|+|P_2|-|P_1 \cap P_2|)$. We first studied the average of *u* (regardless of *D* and *n*) for all pairs of viable metabolisms sampled, and did so for each of the 10 carbon sources. Figure S9 shows that *u* ranges from 0.65 to 0.85, indicating that the majority of novel phenotypes accessible to a local neighborhood is unique to that neighborhood, regardless of the carbon source considered. The average *u* for pairs of metabolisms where all reactions are connected to one another is very similar (Figure S9).

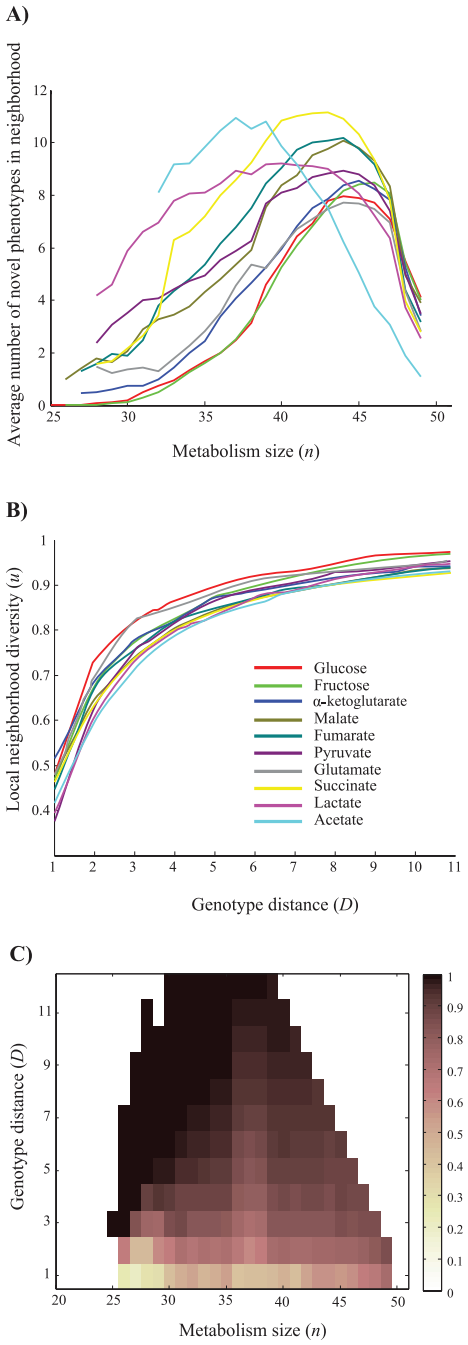



**Figure 4: Neighborhood analysis. a)** Average number of distinct novel metabolic phenotypes in a neighborhood as a function of metabolism size (n) and **b)** average local neighborhood diversity *(u)* as a function of genotypic distance (*D*) (horizontal axes), for metabolisms of size 40 that are viable on different single carbon sources, as indicated in the legend, **c)** Average local neighborhood diversity *u* (see color legend) for metabolism pairs of a given genotype distance (*D, y-axis*) and size (*n, x-axis*), that are viable on glucose, Data are based on 1000 randomly sampled networks for each metabolism size (*n*), genotypic distance (*D*), and carbon source.

Subsequently, we investigated how genotypic distance (*D*), and reaction numbers (*n*) influence phenotypic diversity *u*. We define the genotype distance as the number of reaction changes required to convert one genotype to the other. It is equal to half the Hamming distance of two genotype vectors. Figure 4b illustrates how *u* increases with increasing genotype distance *D*, for metabolisms with *n*=40 reactions required to be viable on each of the 10 carbon sources. The figure illustrates that *u* increases

rapidly until it reaches close to its maximal value at $D \geq 4$. The dependency of $u$ on $D$ is qualitatively identical for different metabolism sizes (Figure S10) and a similar trend is apparent for pairs of metabolisms viable on any one of the 10 carbon sources (Figure S10). Figure 4c shows how $u$ depends on both $D$ and $n$ for metabolisms viable on glucose as the sole carbon source. The figure indicates that regardless of $n$, the fraction of unique phenotypes $u$ increases rapidly with increasing $D$ and reaches a value close to its maximum of $u=1$ at modest $D$. While this qualitative pattern is similar for different carbon sources and reaction numbers (Figure S10), the quantitative relationship between $u$, $D$, and $n$ depends on the carbon source (compare figure 4c based on metabolisms viable on glucose with figure S11 for metabolisms viable on fumarate). Despite such quantitative differences, however, a simple general pattern emerges: except for small metabolisms that are very similar to one another, the majority of phenotypes accessible from any one neighborhood are unique ($u>0.5$). Accessible new phenotypes strongly depend on a metabolism's location in genotype space. Figure S12 shows that for every carbon source, local neighborhood diversity $u$ is greater than expected by chance based on a simple randomization test.

**Genotype networks of different phenotypes are close together in genotype space**

A complementary perspective on the accessibility of novel phenotypes regards the minimal distance of the genotype networks $GN_1$ and $GN_2$ of different phenotypes $P_1$ and $P_2$, i.e., $D_{min}=min\{D(G_1,G_2)|G_1 \in GN_1 \wedge G_2 \in GN_2\}$ where $D(G_1,G_2)$ indicates the genotype distance of two metabolic genotypes $G_1$ and $G_2$. This distance is equivalent to the minimal number of reaction changes that are necessary to convert metabolisms with one phenotype into metabolisms with the other phenotype. We analyzed the distribution of this distance for genotype networks of different phenotypes, focusing our analysis on the giant component for the minority of phenotypes whose viable set of genotypes had more than one connected component (Tables S2, S5, S6, S7 and S8). In analogy to neighboring genotypes, we call two genotype networks neighbors if their minimal distance is $D_{min}=1$. Before embarking on the analysis, we note that only 84 of the $2^{10}=1024$ possible phenotypes for our 10 carbon sources have a genotype associated with it (Figure S13). To see why, consider that all metabolisms viable on fructose are also viable on glucose, because glucose and fructose are biochemically similar and enter central carbon metabolism near one another.

Therefore, there exists no metabolism viable on any combination of carbon sources that contain fructose and lack glucose ($2^8$=256 such "forbidden" phenotypes). Analogous dependencies among other combinations of carbon sources shrink the total number of allowed phenotypes from 1024 to 84.

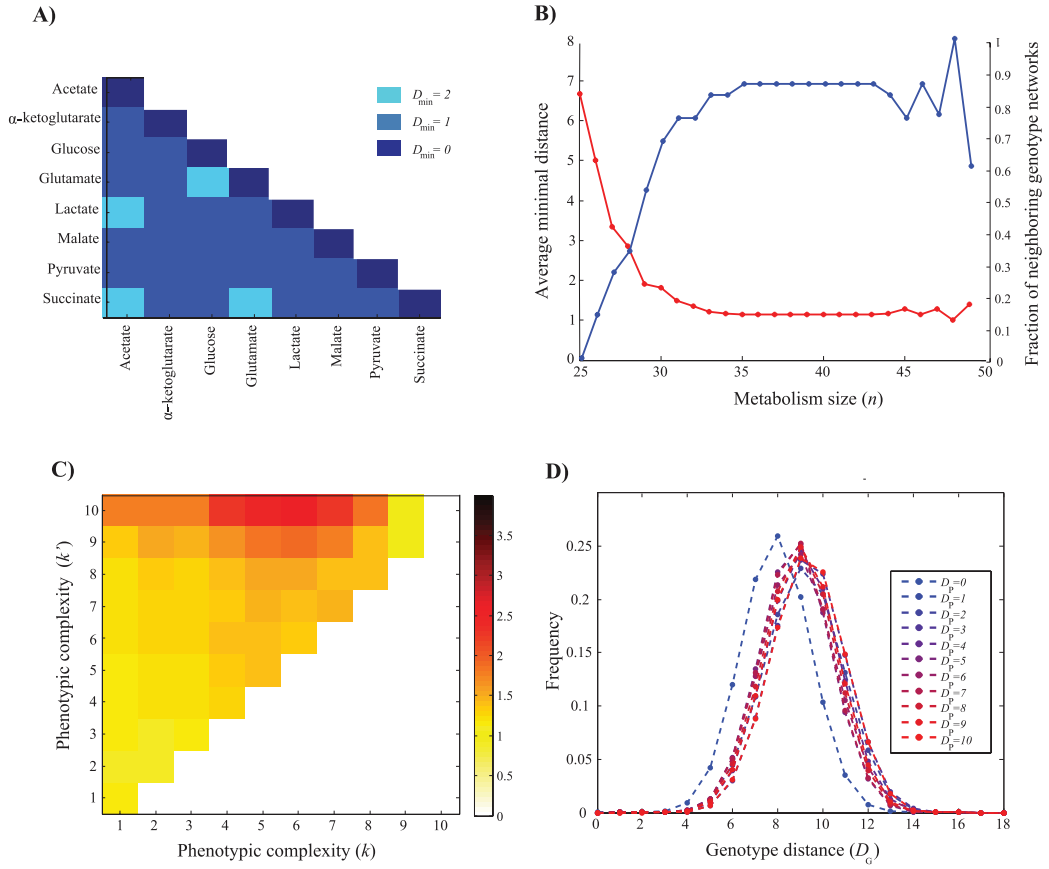



**Figure 5: Minimal distance between genotype networks.** Each rectangular colored area shows the color-coded minimal genotype distance $D_{min}$ between genotype networks comprised of metabolisms that are viable only on carbon sources indicated on the corresponding *(x, y)* positions where (*n=35*). **b)** Average minimal distance among all pairs of genotype networks (red curve, inner vertical axis) and fraction of pairs of genotype networks that are neighbors (blue curve, outer vertical axis) as a function of metabolism size (*n*). **c)** Each rectangular area shows the color-coded average of the minimal distance $D_{min}$ (see color legend) between pairs of genotype networks with phenotypic complexity (*k*, x-axis) and (*k'*, y-axis) for metabolisms of size n=35. At this size, the highest average minimal distances exists for metabolisms of complexity (k, k')=(10, 7), which show $D_{min}$=2.6. **d)** Data points of a given shading indicate the distribution of genotypic distance $D_G$ (x-axis) among pairs of metabolisms with a given phenotypic distance $D_P$, as indicated in the color legend for metabolisms of size *n=30*.

We first analyzed minimal genotype distances for metabolisms viable on only a single carbon source. (Instead of ten carbon sources, there are only eight to consider in this analysis, because all metabolisms viable on fructose are also viable on glucose, and the same holds for malate and fumarate.) Figure 5a shows the minimal distance between such genotype networks at intermediate reaction numbers $n=35$. All except 4 pairs of genotype networks have a minimal distance of one. As $n$ approaches the minimally admissible size for viability, this minimal distance gets modestly larger. However, even at the lowest size ($n=30$) where viable metabolisms exist for all carbon sources, genotype networks are immediately adjacent to one another for 19 out of the 28 possible pairs (Figure S14). Only for a single pair of carbon sources (lactate and acetate) does the minimal distance have the largest value of $D_{min}=6$ (See figure S14, as well as figure S15 for two representative examples of how minimal genotype distances change with increasing metabolism size.)

Figure 5b (red curve) shows the average $D_{min}$ between genotype networks as a function of metabolism size $n$, where the average is over all pairs of carbon sources. This average $D_{min}$ decreases rapidly as $n$ increases. Figure 5b (blue curve) shows the fraction of genotype network pairs for which $D_{min}=1$ increases rapidly with increasing n. In sum, except for the smallest metabolisms, most genotype networks associated with viability on single carbon sources are neighbors and are thus easily reached from one another through single reaction changes.

Next, we extended this analysis to genotype networks for metabolisms viable on $k>1$ carbon sources. For brevity, we refer to $k$ as the phenotypic complexity of a metabolism. Figure 5c shows the average minimal distance, i.e., the average number of reaction changes minimally needed to reach a genotype network with phenotypic complexity $k$ from a network with complexity $k'$, where $k$ and $k'$ range between 1 and 10 and where $n$=35. The figure demonstrates that $D_{min}$ generally increases with either $k$ or $k'$, and that it is most difficult to reach metabolisms of intermediate phenotypic complexity ($k=7$) from metabolisms with the highest complexity ($k=10$), where $D_{min} \approx 2.60$. At the same time, $D_{min}$ decreases with increasing metabolism size (compare figures S16a, 5c, and S16b). A complementary analysis in figure S17, which focuses on the fraction of neighboring genotype networks ($D_{min}=1$) as a function of phenotypic complexity, shows a similar pattern: At any one metabolism size, the fraction of neighboring genotype networks decreases with phenotypic complexity.

Figures S18 and S19 show that these patterns also hold if we consider only those metabolisms where all reactions are connected to one another. Finally, figures S20 and S21 show that except in the smallest metabolisms, the average of the minimal distance among all pairs of genotype networks is close to 1 (See supplementary Text S4) indicating that most genotype networks abut each other in the genotype space.

**Genotypic similarity versus phenotypic similarity**

In our last analysis, we asked how many reaction changes are required to change a metabolism with a given genotype $G_1$ and phenotype $P_1$ into a metabolism $G_2$ with an arbitrary new phenotype $P_2$. More specifically, we were interested in how the (Hamming) distance between the two genotypes $D_G$ depends on the (Hamming) distance $D_P$ between the phenotypes $P_1$ and $P_2$. $D_P$ is based on the representation of phenotypes as ten-dimensional binary vectors indicating viability on each of our ten carbon sources. The answer can indicate how difficult it is to reach a distant phenotype from any one viable genotype. To compute this distance, we first sampled 1000 metabolisms with a given size ($n=35$) among all viable metabolisms, regardless of the number of carbon sources that they are viable on. Next, for each sampled metabolism, we exhaustively calculated its genotype distance and its phenotype distance against all viable metabolisms of the same size.

The results are shown in figure 5d. The left-most distribution shows the genotype distances of those genotypes that have identical phenotypes ($D_P=0$), for comparison against those genotypes that have different phenotypes ($D_P>0$). The figure shows that the distributions of genotypic distance of phenotypes with varying phenotypic distance are very similar. Moreover, the mean distance is shifted only slightly to the right (by one reaction change) relative to the mean distance of genotypes with identical phenotypes. Similar patterns exist for metabolisms of different sizes (Figure S22) and for metabolisms consisting only of connected reactions (Figure S23). Taken together, these patterns imply that reaching a genotype with an arbitrarily distant phenotype from any one point in genotype space does not require many more reaction changes than traversing a genotype network.

## 6.4. Discussion

Any population of organisms evolves in a space of possible genotypes and their phenotypes. Such a "space of the possible" may harbor new and useful phenotypes, but it may also constrain a population's potential for innovation. To understand both innovation opportunities and constraints, it is necessary to understand the organization of such spaces. For a population of organisms whose metabolic reaction network changes through addition and elimination of individual reactions, the relevant space is a space of metabolic genotypes – each genotype represents a specific set of reactions – and the phenotypes they encode. If one considers all known biochemical reactions, this space is vast, around $10^{2000}$ genotypes [54–56]. Past analyses thus relied on sampling [11–18]. We here complement sampling-based analyses through an exhaustive enumeration of all $2.25 \times 10^{15}$ genotypes in the subspace whose reactions are involved in central carbon metabolism.

Our most basic observation regards the fraction of viable metabolisms, which is tiny, ranging from $10^{-8}$ (for acetate) to $10^{-6}$ (for glucose) for metabolisms viable on at least one carbon source, and becomes substantially smaller for metabolisms viable on all 10 carbon sources ($10^{-10}$). However, because the entire space is large, these tiny fractions translate into sizable numbers of 10,850,304 and 1,549,771,520 viable metabolisms on acetate and glucose respectively and 1,029,375 metabolisms on all 10 carbon sources. Because we observe that most viable metabolisms do not contain any reactions disconnected from the rest of the reaction network, most of these variant metabolisms are not trivially obtainable by addition of disconnected reactions to a functional metabolic core. Thus, even in the modest genotype space created by all subsets of 51 biochemical reactions, there are myriad alternative metabolic organizations that achieve viability through different means.

The fraction of viable metabolisms is not uniformly distributed among metabolisms with different complexity, i.e., number $n$ of reactions. Regardless of the specific phenotype one considers, the fraction has a minimum at the minimal complexity needed for viability (the smallest $n_{min}=23$ exists for viability on glucose, and the largest $n_{min}=30$ for metabolisms viable on acetate) and increases rapidly towards a fraction of one for the largest metabolisms. In addition, metabolisms with the same phenotype form large connected networks. These networks may contain more than one connected components at the smallest complexity but they congeal to a single

component at intermediate complexity. Moreover, this "giant" component [55,58] extends increasingly far through genotype space until its diameter becomes equal to the entire space at intermediate reaction numbers. Taken together, this means that central carbon metabolism shows substantial internal flexibility in its organization. It can be altered one reaction at a time to create different metabolic architectures with the same phenotype, exemplified by metabolisms like those shown in figure 2 that differ in about half of their reactions. The exceptions to this rule can be found among the smallest, least complex metabolisms. Evolutionary change that alters metabolic genotypes without altering phenotypes is easier in complex metabolisms with many reactions.

A metabolism's complexity is also relevant for its potential to encounter novel metabolic phenotypes in its immediate neighborhood, i.e., through single reaction changes. The number of different novel phenotypes that are encountered in a metabolism's neighborhood exceeds one for all but the smallest metabolisms, and it rises to a maximum of 8-11 at intermediate metabolic complexity. That it does not increase further for larger $n$ is the result of a model limitation, namely that we consider only ten carbon sources. Large metabolisms are already viable on most of these carbon sources, such that further addition of reactions can no longer create novel phenotypes that are viable on additional carbon sources (Figure S24). These observations suggest that the number of novel phenotypes accessible through single reaction changes increases with metabolic complexity.

While the neighborhoods of most metabolisms contain multiple novel phenotypes, the identity of these phenotypes depends strongly on a metabolism's location in genotype space. That is, the majority of novel phenotypes contained in the neighborhoods of two closely related metabolisms are not shared between these neighborhoods. In other words, any one novel phenotype tends to occur either in one or the other neighborhoods, but not in both. This means that the evolutionary potential of any one metabolism with a given phenotype is contingent upon its genotype. This contingency is alleviated by the connectedness of different metabolisms with the same phenotype. Because their genotype can be altered without phenotypic change, phenotype-preserving evolutionary change in genotypes can make different neighborhoods and their novel phenotypes accessible. In this regard, it is also relevant that the minimal distance of most genotype networks is small, such

that one can reach novel phenotypes through one or few genotypic changes from any one genotype network. The exceptions to this rule come again from the smallest metabolisms, suggesting that metabolic complexity also facilitates this aspect of phenotypic evolution.

One major obstacle to genotype-phenotype mapping comes from the vast size of genotype spaces. We could overcome this obstacle here by considering a modestly sized genotype space of $10^{15}$ carbon metabolism variants. Carbon metabolism is an attractive small study system, because it plays a key role in extracting energy from extracellular carbon sources. In addition, we limited ourselves to viability on 10 different carbon sources, and thus to potentially $2^{10}$=1024 viability phenotypes. This number is sufficiently small to be computationally tractable, yet large enough to allow quantitative analyses, for example about the phenotypic diversity of different neighborhoods. Even so, we could enumerate the phenotype of all $10^{15}$ metabolisms only after taking advantage of certain relationships among metabolisms, such as that the children of inviable metabolisms are also inviable.

A limitation of analyzing central carbon metabolism is that it is not suited to study metabolic innovation in essential nutrients like nitrogen or sulfur. It thus remains to be seen whether similar principles hold for these nutrients and their metabolism. Sampling studies of larger genotype spaces suggest that this is indeed the case [14,59].

Another limitation of our analysis is its focus on evolutionary constraints that are imposed only by the presence or absence of reactions. In other words, we have neglected regulatory constraints that can arise through suboptimal expression or regulation of an enzyme. In this regard, we note that such constraints are most important if one focuses on the quantitative predictions of biomass growth via flux balance analysis [60]. In contrast, we here focus on the purely qualitative prediction of viability, i.e., whether biomass can be produced at all. This qualitative phenotype is biologically relevant if one considers that many organisms grow slowly in their native environment [61–63]. In addition, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments [60,64,65].

A third limitation comes from the fact that the exhaustive enumeration we pursue requires us to start from a limited "universe" of chemical reactions. The choice of reactions in this universe may introduce some biases into our analysis. For example, it is not clear whether the small number of metabolisms viable on acetate is a result of this choice, or whether it would also persist in an unbiased analysis of larger, genome scale metabolisms. However, we note that our core results agree well with previous studies based on sampling of genome-scale metabolisms that comprise more than a thousand reactions. For instance, genome-scale metabolisms with the same phenotype can be Genotypically very different [12] and usually form single connected genotype networks [19], which is in line with our present observation that genotype networks have large diameter. Moreover, genome-scale metabolisms can encounter many new phenotypes in their immediate neighborhood, and the neighborhood of different genome-scale metabolisms contains different novel phenotypes [12,14,59], which is consistent with our neighborhood analysis. Finally, genome-scale metabolisms with different phenotypes can be found close together in genotype space [12,59].

A final limitation comes from our definition of a metabolic genotype centered on individual metabolic reactions. Although widely used in the field [44, 66-70] this notion of a genotype does not take into account that some reactions are catalyzed by multiple enzymes [71], and conversely, that some enzymes catalyze multiple reactions [72-74]. Our focus on the metabolic reaction as the most elementary unit of evolutionary change should not distract from the fact that actual change in metabolic systems may be more complex.

In sum, our exhaustive analysis of central carbon metabolism's genotype space reveals an organization that is conducive to both the preservation of phenotypes in the face of genotypic change, and the exploration of new phenotypes. Because metabolisms with the same phenotype can be connected to one another through single reaction changes, viability phenotypes can be preserved through substantial genotypic change. Because connected sets of genotypes associated with different phenotypes are close to each other in genotype space, novel phenotypes can often be reached with few or no transitions through intermediate phenotypes. These principles only break down in metabolisms with low complexity, close to the minimal number of reactions needed for viability. Thus, increasing metabolic complexity enhances both the potential for evolutionary conservation and innovation.

## 6.5. Methods

### Flux balance analysis

Flux balance analysis (FBA) is a constraint-based computational modeling approach that is widely used for quantitative analysis and modeling of metabolism. FBA predicts the metabolic flux through every reaction in a metabolism, based on information about the metabolism's reactions, as well as about the stoichiometric coefficients of the reactants in each reaction. These coefficients are contained in the stoichiometric matrix $S$, which is of dimension $m \times n$, where $m$ and $n$, respectively, denote the number of metabolites and the number of reactions in the metabolism. An important assumption behind FBA is that the concentration of metabolites does not change, that is, the metabolism is in a steady state. This assumption imposes mass conservation constraints on the metabolites in the network, which can be mathematically expressed as $Sv = 0$, where $v$ is the vector of metabolic fluxes $v_i$ through reaction $i$. The possible solutions of the above equation are the allowable flux vectors, which form the null space of the stoichiometric matrix $S$. This space is further constrained by the fact that each reaction has a maximally and minimally possible flux through it. FBA uses an optimization technique called linear programming to identify among the allowed flux vectors those vectors that maximize an objective function $Z$. This task can be formulated as finding a flux vector $v^*$ with the property

$$v^* = max_v\, Z(v) = \{c^T v | S.v = 0,\ a \leq v \leq b\}$$

where the vector $c$ is a set of scalar coefficients representing the maximization criterion, and each entry $a_i$ and $b_i$ of vectors $a$ and $b$, respectively, indicates the minimally and maximally possible flux through reaction $i$.

We are interested in knowing whether a metabolic reaction network can sustain life in a given environment, that is, whether it can synthesize all essential small biomass molecules required for survival and growth. In this work, we used 13 well-known precursor substances from central carbon metabolism as the set of required biomass molecules (table S9). As is common in FBA applications [75–79], the objective function that we use for the linear programming of FBA is a biomass reaction that transforms 13 precursors into biomass (table S9). We used the package CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve linear programming problems.

**Chemical environments**

In addition to a stoichiometric matrix and an objective function (biomass growth), it is necessary to define the chemical composition of an environment that contains different nutrients required for the synthesis of biomass precursors. We consider minimal growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron ($Fe^{2+}$ and $Fe^{3+}$), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc. All these nutrients except the carbon source are shared between different minimal environments. Each minimal environment contains a different one of the 10 carbon sources acetate, $\alpha$-ketoglutarate, fumarate, fructose, glucose, glutamate, lactate, malate, pyruvate, and succinate in our analysis.

**Reaction universe**

The "universe" of reactions in our metabolic genotype space is based on *E. coli* central carbon metabolism [80], from which we deleted four reactions involved in ethanol synthesis, metabolism, and transport. We also grouped the reactions catalyzed by aconitase A and aconitase B into one reaction, to render exploration of all metabolisms that consist of different combinations of these reactions feasible. The final reaction set consists of $N = 51$ intracellular reactions that can be present or absent in different metabolisms (table S9). Twenty different transport reactions, which are necessary to import nutrients or excrete waste products, are present in all metabolisms we study, i.e., we do not vary their presence among metabolisms.

**Metabolic genotypes, metabolic phenotypes and genotype networks**

The nucleotide sequence of the genes encoding the enzymes catalyzing a metabolism's reactions constitutes the metabolic genotype of an organism. However, for our purpose, we use a more compact representation of a metabolic genotype, in which we represent this genotype as a binary vector whose *i*-th entry corresponds to the *i*-th reaction in our reaction universe. The *i*-th entry is equal to one if an organism's genome encodes an enzyme capable of catalyzing this reaction, and zero

otherwise. The genotype space of all possible metabolisms comprises $2^N$ metabolisms ($N=51$). Each metabolism can be thought of as a point in this space. We call metabolisms that are able to synthesize all 13 biomass precursors from nutrients in this environment *viable.* More precisely, we consider a metabolism viable on a carbon source if its biomass synthesis rate is greater than one percent of the biomass synthesis rate of the network formed by all $N = 51$ reactions on that carbon source [19]. Our definition of a carbon utilization phenotype for any one genotype is based on its viability on different carbon sources. Specifically, we assign to each genotype a phenotype vector of length 10, equal to the number of distinct carbon sources we consider. The *i*-th entry of the phenotype vector corresponds to the minimal environment containing the *i*-th sole carbon source. This entry equals 1 if the metabolism is viable on that minimal environment and zero otherwise. In other words, we consider $2^{10}$=1024 distinct metabolic phenotypes. We partition metabolic genotype space into distinct sets of genotypes, each with a different phenotype. Each such genotype set can be further partitioned into subsets of metabolisms with different sizes, that is, different numbers *n* of chemical reactions. If a subset of metabolisms (genotypes) forms a connected graph [57,58] in genotype space, we call that graph a genotype network. For some analyses, it is useful to consider a modified definition of a phenotype that just specifies whether a metabolism is viable on *at least* a specific set of carbon sources – it may be viable on other carbon sources as well. With this phenotype definition, different genotype sets and genotype networks can overlap.

**Exhaustive enumeration of viable metabolisms**

To exhaustively characterize the phenotype of every single one among $2^{51}$ ($10^{15}$) metabolic genotypes one would need to use FBA $10^{15}$ times. Given that a typical FBA computation takes of the order of $10^{-2}$ seconds of CPU time, exhaustive computational phenotyping would require $10^5$ years and would thus be prohibitive. However, one can take advantage of two simple facts to render this computation feasible in approximately 10 days [19,81]. First, six among the 51 internal reactions of central carbon metabolism are essential for viability on every carbon source we consider [15]. The corresponding entries of the genotype vector need to be set to one,

which reduces the number of required FBA computations by a factor $2^6$ from $2^{51}$ ($10^{15}$) to $2^{45}$ ($10^{13}$).

Second, all metabolisms ("children") that contain a subset of the reactions of an inviable metabolism ("parent") will also be inviable, because deleting one or more reactions from an inviable metabolism cannot result in a viable metabolism. In an earlier work, one of us designed an algorithm to take advantage of this observation [81]. It divides genotype vectors of length 45 into 5 distinct sub-vectors of length 9, determines all viable genotypes originating from each binary sub-vector of length 9 ($2^9$=512 subvectors), and merges the sub-vectors of viable genotypes in a five-step procedure to determine all the viable genotypes on a given carbon source. At each step, only sub-vectors that preserve viability are merged, which dramatically decreases the number of required FBA tests to approximately$10^9$ total tests.

Once the set $GN(C_i)$ of metabolisms viable on each of the 10 carbon sources $C_i$ has been determined, one can easily identify the set of metabolisms that are viable on a given subset $S$ of the 10 carbon sources. Specifically, $V(S)=\{G\in\Omega,\ \forall\ C_i\in S|G\in GN(C_i)\}$ where $G$ denotes a given genotype belonging to the genotype space $\Omega$. Similarly, one can define the set of genotypes that are exclusively viable on carbon sources in $S$ as: $V(S)=\{G\in\Omega,\ \forall\ C_i\in S,\ \forall\ C_j\in S' |G\in GN(C_i),\ G\notin GN(C_j)\}$ , where $S'$ denotes the complement of $S$.

**Connectedness of metabolic genotype networks**

We can represent each set of metabolisms (genotypes) of a given size $n$ that is viable on a subset $S$ of carbon sources as a graph. The nodes of this graph are metabolisms. Two viable metabolisms $A$ and $B$ are connected by an edge if metabolism $A$ is convertible to $B$ via a reaction swap, that is, by deleting a reaction that $A$ possesses but $B$ lacks, followed by adding a reaction that $B$ possesses but $A$ lacks. We note that such a swap leaves the number of reactions constant, as is required for metabolisms that have the same size. However, any one reaction swap can be decomposed into an addition of a reaction followed by a deletion of a reaction. In other words, viable metabolisms that are neighbors based on reaction swaps are also connected through single reaction changes.

If this graph of metabolisms is connected, then every single metabolism in it can be reached from any other metabolism via a sequence of single reaction changes. Otherwise, the graph fragments into several disconnected components, one of which may be much larger than the others, and is therefore often also called the "giant" component [57,58]. The connected components of a graph can be computed with the aid of algorithms like Breadth-First Search (BFS) [82]. BFS requires a graph's adjacency list, which contains the neighbors of each node in the graph. To generate this list for a genotype network, one needs to compare $V^2$ pair of metabolisms to ascertain whether they are neighbors. Because the genotype networks we consider may comprise many thousands to millions of metabolisms, doing so would be computationally prohibitive. Moreover, storing an entire adjacency list causes memory problems in large genotype networks. Therefore, we developed an algorithm to examine connectedness of genotype networks [81], which differs from conventional BFS in that (i) it does not need to fill the adjacency matrix in advance, and (ii) it can avoid comparing genotypes that could not possibly be neighbors. In doing so, it reduces the number of genotype comparisons from $V^2$ to $mV$ where $m$ is the average number of a genotype's neighbors. Using this algorithm, we could determine connectedness of genotype networks comprising as many as $10^6$ metabolisms. For larger metabolisms, where the requirements for storing all metabolisms becomes prohibitive, we could determine genotype network connectivity by taking advantage of the following simple principle: If a genotype network of metabolisms of size $n$ is connected, then the genotype network of metabolisms of size $n+1$, each genotype of which is constructed by adding an additional reaction to each genotype belonging to the genotype networks of size $n$, is also connected [19]. In other words, we could infer the connectedness of larger genotype networks from the connectedness of smaller ones.

**Diameter of metabolic genotype networks**

The diameter of a graph (genotype network) is the maximum length of all shortest paths between any pair of nodes that reside in the same connected component. In a connected genotype network, the shortest path between any pair of genotypes is the minimal number of reaction changes required to convert the two genotypes into each other. This number is equivalent to half of the Hamming distance between the binary

vectors representing the genotypes. To determine the diameter of a genotype network, we needed to identify those genotypes whose Hamming distance is maximal. We were able to do this through exhaustive enumeration for genotype networks with fewer than $10^5$ metabolisms. For larger genotype networks, we could only determine a lower bound on the diameter, and we did so by sampling $10^5$ genotypes from a given genotype network and determining the two genotypes with the largest distance among them. We note that maximum diameter of genotype space as a function of metabolism size ($n$) is *Min {(51-n), n},* and in most of the large genotype networks, the sampling based diameter estimate was equal to the maximum diameter of the genotype space. This confirms that the sample size was big enough to accurately estimate the diameter of the genotype networks.

## 6.6. References


1. Correns C. (1900) G. Mendel's law on the behaviour of progeny of variable hybrids. Ber Dtsch Bot Ges: 8: 156–168.
2. Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM (2000) An Introduction to Genetic Analysis. New York: W. H. Freeman.
3. Palsson BØ (2006) Systems Biology: Properties of Reconstructed Networks. Cambridge: Cambridge University Press.
4. Wagner A (2012) Metabolic networks and their evolution. Adv Exp Med Biol 751: 29–52.
5. Kaya H, Shimizu S (2002) Computational methods in protein folding: Scaling a hierarchy of complexities. In Jiang T, Xu Y, Zhang MQ, editors. Current Topics in Computational Molecular Biology. Cambridge, Massachussetes, USA. pp. 403–447.
6. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, et al. (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie Chem Mon 125: 167–188.
7. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protoc 2: 727–738.
8. Irbäck A, Troein C (2002) Enumerating Designing Sequences in the HP Model. J Biol Phys 28: 1–15.
9. Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, et al. (1996) Analysis of RNA sequence structure maps by exhaustive enumeration .1. Neutral networks. Monatshefte für Chemie 127: 374.
10. Nochomovitz YD, Li H (2006) Highly designable phenotypes and mutational buffers emerge from a systematic mapping between network topology and dynamic output. Proc Natl Acad Sci U S A 103: 4180–4185.

11. Samal A, Matias Rodrigues JF, Jost J, Martin OC, Wagner A (2010) Genotype networks in metabolic reaction spaces. BMC Syst Biol 4: 30.

12. Matias Rodrigues JF, Wagner A (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput Biol 5: e1000613.

13. Samal A, Wagner A, Martin OC (2011) Environmental versatility promotes modularity in genome-scale metabolic networks. BMC Syst Biol 5: 135.

14. Matias Rodrigues JF, Wagner A (2011) Genotype networks, innovation, and robustness in sulfur metabolism. BMC Syst Biol 5: 39.

15. Barve A, Rodrigues JFM, Wagner A (2012) Superessential reactions in metabolic networks. Proc Natl Acad Sci U S A 109: E1121–30.

16. Barve A, Wagner A (2013) A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500: 203–206.

17. Bilgin T, Wagner A (2012) Design constraints on a synthetic metabolism. PLoS One 7: e39903.

18. Ebenhöh O, Heinrich R (2003) Stoichiometric design of metabolic networks: multifunctionality, clusters, optimization, weak and strong robustness. Bull Math Biol 65: 323–357.

19. Barve A, Hosseini S-R, Martin OC, Wagner A (2014) Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. BMC Syst Biol 8: 48.

20.. Cline RE, Hill RH, Phillips DL, Needham LL (n.d.) Pentachlorophenol measurements in body fluids of people in log homes and workplaces. Arch Environ Contam Toxicol 18: 475–481.

21. Copley SD (2009) Evolution of efficient pathways for degradation of anthropogenic chemicals. Nat Chem Biol 5: 559–566.

22. Rehmann L, Daugulis AJ (2008) Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. Biotechnol Bioeng 99: 521–528.

23. Van der Meer JR, Werlen C, Nishino S, Spain J (1998) Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. Appl Environ Microbiol 64: 4185–4193.

24. Dantas G, Sommer MOA, Oluwasegun RD, Church GM (2008) Bacteria subsisting on antibiotics. Science 320: 100–103.

25. Postgate JR (1994) The Outer Reaches Life. Cambridge: Cambridge University Press.

26. Detkova EN, Boltyanskaya Y V. (2007) Osmoadaptation of haloalkaliphilic bacteria: Role of osmoregulators and their possible practical application. Microbiology 76: 511–522.

27. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.

28. Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary Origins of Genomic Repertoires in Bacteria. PLoS Biol 3: e130.

29. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3: 711–721.

30.. Papagianni M (2012) Recent advances in engineering the central carbon metabolism of industrially important bacteria. Microb Cell Fact 11: 50.

31. Romano AH, Conway T (n.d.) Evolution of carbohydrate metabolic pathways. Res Microbiol 147: 448–455.

32. Meléndez-Hevia E, Waddell TG, Heinrich R, Montero F (1997) Theoretical approaches to the evolutionary optimization of glycolysis--chemical analysis. Eur J Biochem 244: 527–543.

33. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. Proc Natl Acad Sci U S A 110: 10039–10044.

34. De Rosa M, Gambacorta A, Nicolaus B, Giardina P, Poerio E, et al. (1984) Glucose metabolism in the extreme thermoacidophilic archaebacterium Sulfolobus solfataricus. Biochem J 224: 407–414.

35. Danson MJ (1989) Central metabolism of the archaebacteria: an overview. Can J Microbiol 35: 58–64.

36. Bar-Even A, Flamholz A, Noor E, Milo R (2012) Rethinking glycolysis: on the biochemical logic of metabolic pathways. Nat Chem Biol 8: 509–517.

37. Huynen MA, Dandekar T, Bork P (1999) Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol 7: 281–291.

38. Fuhrer T, Fischer E, Sauer U (2005) Experimental identification and quantification of glucose metabolism in seven bacterial species. J Bacteriol 187: 1581–1590.

39. Noor E, Eden E, Milo R, Alon U (2010) Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. Mol Cell 39: 809–820.

40. Price ND, Reed JL, Palsson BØ (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2: 886–897.

41. Edwards JS, Covert M, Palsson B (2002) Metabolic modelling of microbes: the flux-balance approach. Environ Microbiol 4: 133–140.

42. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28: 245–248.

43. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3: 121.

44. Edwards JS, Palsson BO (2000) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A 97: 5528–5533.

45. Edwards JS, Ibarra RU, Palsson BO (2001) In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19: 125–130.

46. Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A 99: 15112–15117.

47. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res 13: 244–253.

48. Wang Z, Zhang J (2009) Abundant indispensable redundancies in cellular metabolic networks. Genome Biol Evol 1: 23–33.

49. Papp B, Pál C, Hurst LD (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429: 661–664.

50. Schuetz R, Kuepfer L, Sauer U (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol 3: 119.

51. Bonarius HPJ, Schmid G, Tramper J (1997) Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends Biotechnol 15: 308–314.

52. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U (2012) Multidimensional optimality of microbial metabolism. Science 336: 601–604.

53. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301–312.

54. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M (2002) LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res 30: 402–404.

55. Goto S, Nishioka T, Kanehisa M (2000) LIGAND: chemical database of enzyme reactions. Nucleic Acids Res 28: 380–382.

56. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–60.

57. Bollobás B (2001) Random Graphs. Cambridge: Cambridge University Press.

58. Newman M (2010) Networks: An Introduction. Oxford: Oxford University Press.

59. Wagner A, Andriasyan V, Barve A (2014) The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. J Mol Biochem 3.

60. Ibarra RU, Edwards JS, Palsson BO (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420: 186–189.

61. Vieira-Silva S, Rocha EPC (2010) The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet 6: e1000808.

62. Cox RA (2004) Quantitative relationships for specific growth rates and macromolecular compositions of Mycobacterium tuberculosis, Streptomyces coelicolor A3(2) and Escherichia coli B/r: an integrative theoretical approach. Microbiology 150: 1413–1426.

63. Kirschner D, Marino S (2005) Mycobacterium tuberculosis as viewed through a computer. Trends Microbiol 13: 206–211.

64. Fong SS, Palsson BØ (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat Genet 36: 1056–1058.

65. Fong SS, Marciniak JY, Palsson BO (2003) Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J Bacteriol 185: 6400–6408.

66. Edwards J.S., Palsson B.Ø. (1999) Systems Properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274:17410–17416.

67. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions usingEscherichia coli. Nat Biotechnol 26: 659–667.

68. Oberhardt MA, Palsson BO, Papin JA (2009) Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5: 320

69. McCloskey D, Palsson BØ, Feist AM (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Molecular Systems Biology 9: 1–15.

70. Lewis NE, Nagarajan H, Palsson BO (2012) Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat Rev Microbiol 10: 291–305.

71. Hunter, R. L. and C.L. Merkert. (1957) Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science* 125: 1294-1295.

72. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. Annu Rev Biochem 79:471–505. 45.

73. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD (2010) Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5′-phosphate synthesis. Mol Syst Biol 6:436.

74. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO: Network context and selection in the evolution to enzyme specificity**.** Science (New York, NY) 2012, **337:**1101-1104.

75. Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26: 659–667.

76. Alper H, Miyaoku K, Stephanopoulos G (2005) Construction of lycopene-overproducing *E. coli* strains by combining systematic and combinatorial gene knockout targets. Nat Biotechnol 23: 612–616.

77. Pharkya P, Burgard AP, Maranas CD (2004) OptStrain: a computational framework for redesign of microbial production systems. Genome Res 14: 2367–2376.

78. Lee D-S, Burd H, Liu J, Almaas E, Wiest O, et al. (2009) Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. J Bacteriol 191: 4015–4024.

79. Jamshidi N, Palsson BØ (2007) Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. BMC Syst Biol 1: 26.

80. Orth JD, Fleming RMT, Palsson BØ (2010) Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. EcoSal Plus 1.

81. Hosseini S-R (2013) Exhaustive genotype-phenotype mapping in metabolic genotype space. M.Sc. Thesis, Swiss Federal Institue of Technology Zürich. Available: http://ecollection.library.ethz.ch/view/eth:7522?q=(keywords_en:PHENOTYPE. Accessed 14 November 2013.

82. Hopcroft J, Tarjan R (1973) Algorithm 447: efficient algorithms for graph manipulation. Commun ACM 16: 372–378.

## 6.7. Supplementary Information

### Text S1: Size differences between minimal metabolisms viable on acetate and glucose

As we mentioned in the main text, the minimal number of reactions in a viable metabolism (i.e. minimal metabolism) is not the same for different carbon sources. For example, it varies from $n_{min}=23$ for glucose and fructose to $n_{min}=30$ for acetate (Figure S2).The reason why a minimal metabolism using acetate needs more reactions is that it uses reactions from gluconeogenesis (Figure S2b) whereas glucose metabolism relies on glycolysis (Figure S2a). More specifically, the acetate (ac) to be metabolized is converted to acetyl-coenzyme A (accoa, Figure S2b) through two consecutive reactions, catalyzed respectively by acetate kinase and phosphotransacetylase. Acetyl-coenzyme A is then used in the glyoxylate cycle in order to produce malate and oxaloacetate. The latter is subsequently converted to phosphoenolpyruvate by phosphoenolpyruvate carboxylase. The gluconeogenic pathway that ensures production of the remaining essential biomass molecules is initiated from phosphoenolpyruvate. Thus, the two reactions in the production of acetyl-coenzyme A, plus the five reactions needed to produce phosphoenolpyruvate through the glyoxylate cycle (i.e. isocitrate lyase, malate synthase, succinate dehydrogenase, fumarase, and malate dehydrogenase) account for the seven additional reactions that are required for viability of a minimal metabolism on acetate. In contrast, glucose neither needs the reactions of the glyoxylate shunt nor those of acetate metabolism, and can thus be metabolized with fewer reactions.

### Text S2: Prediction of the number of viable metabolisms based on binomial coefficients

Figures S3a and S3b illustrate that a shifted binomial coefficient (i.e., $\binom{N-n_{\min}}{n-n_{\min}}$) qualitatively predicts the relationship between reaction number and the number of metabolisms viable on a carbon source. However, it overestimates this number, especially for metabolisms at low- and intermediate sizes. The reason for the qualitative agreement stems from the fact that adding reactions to a viable metabolism will not render this metabolism inviable. Assume that only a single viable

metabolism with a set of $n_{\min}$ reactions exists for a given carbon source. Adding any subset of the remaining $N - n_{\min}$ reactions to this metabolism will not render it inviable. To obtain a viable metabolism with size $n$ that lies $n - n_{\min}$ reactions above the minimum, one has $\binom{N-n_{\min}}{n-n_{\min}}$ possible choices of $n - n_{\min}$ reactions, which explains the qualitative binomial relationship.

The discrepancies between the binomial relationship and the data stem from violations of this assumption. In previous contributions [1,2] we showed that there are usually multiple minimal viable metabolisms $v_{\min}$. For instance, there are 3 and 4 minimal viable metabolisms on glucose and acetate, respectively (Table S1). If one extends the above line of reasoning to incorporate this observation, one arrives at the relationship $v_{\min}\binom{N-n_{\min}}{n-n_{\min}}$ as a predictor for the number of viable metabolisms, which is shown in dashed lines in figures S3a an S3b. This predictor is clearly superior to the shifted binomial coefficient, but a slight discrepancy persists.

This discrepancy has two causes, one a source of underestimating, the other a source of overestimating numbers of viable metabolisms. To explain them, we briefly review some previous observations on minimal metabolisms [1,3,4]. By definition, a minimal metabolism is one from which no reaction can be removed without eliminating viability. Importantly, a minimal metabolism is not necessarily the smallest possible viable metabolism, because there may be metabolisms with more than $n_{\min}$ reactions, from which no reactions can be removed. On glucose, for example, the smallest viable metabolisms with $n_{\min} = 23$ reactions is also a minimal metabolism, but there also exist other metabolisms, at sizes *n=24* (8 metabolisms), at *n=25* (23 metabolisms), up through *n=30*, from which no reaction can be removed. To each of these metabolisms, any number of reactions can be added without abolishing viability, and each of them can thus contribute to the number of viable metabolisms at larger sizes. Not taking them into account is a reason why the predictor $v_{\min}\binom{N-n_{\min}}{n-n_{\min}}$ underestimates the total number of viable metabolisms at these sizes.

To understand how the binomial predictor can also overestimate the number of viable metabolisms, consider two minimal metabolisms A and B of the same size, and the metabolism AB consisting of the union of their constituent reactions. Because AB can be viewed as resulting from adding sets of reactions to A or to B, it should be viable ($AB = A \cup (AB \backslash B)$ or $AB = B \cup (AB \backslash A)$). The above binomial expression $V_{min}\binom{N - n_{min}}{n - n_{min}}$ implicitly counts the metabolism AB twice in computing the number of viable metabolisms, whereas AB should only be counted once. This cause is responsible for the overestimation of total metabolism sizes at large *n*, for example at *n=49* to *n=51* for glucose.

**Text S3: The unimodal distribution of the number of novel phenotypes in the neighborhood of viable metabolisms**

To explain the unimodal distribution from figure 4a, consider first metabolisms *M* whose size *n* is below this peak. As we mentioned in the text, the fraction of viable metabolisms increases faster than exponentially with increasing reaction numbers (Figure 1b). At larger sizes, more metabolisms in any one metabolism's neighborhood are thus viable, which also increases the number of novel phenotypes that these metabolisms can have. This observation can explain that the number of accessible novel phenotypes increases with *n*, at least up to intermediate *n*. The fraction of all viable metabolisms (Figure 1b) continues to increase above the value of *n* where the number of accessible novel phenotypes is maximal, but a second pattern becomes important above this peak *n*. Specifically, the total number of distinct phenotypes that all metabolisms of a given size can have is highest for intermediate sizes and decreases at the largest sizes. The reason is that metabolisms *M* containing most reactions also tend to be viable on most carbon sources, such that there are fewer possible phenotype vectors with more ones than *M* (Figure S11). This is why the average number of accessible novel phenotypes declines beyond the peak *n*.

**Text S4: Minimal genotype network distance as a function of phenotypic complexity and metabolism size**

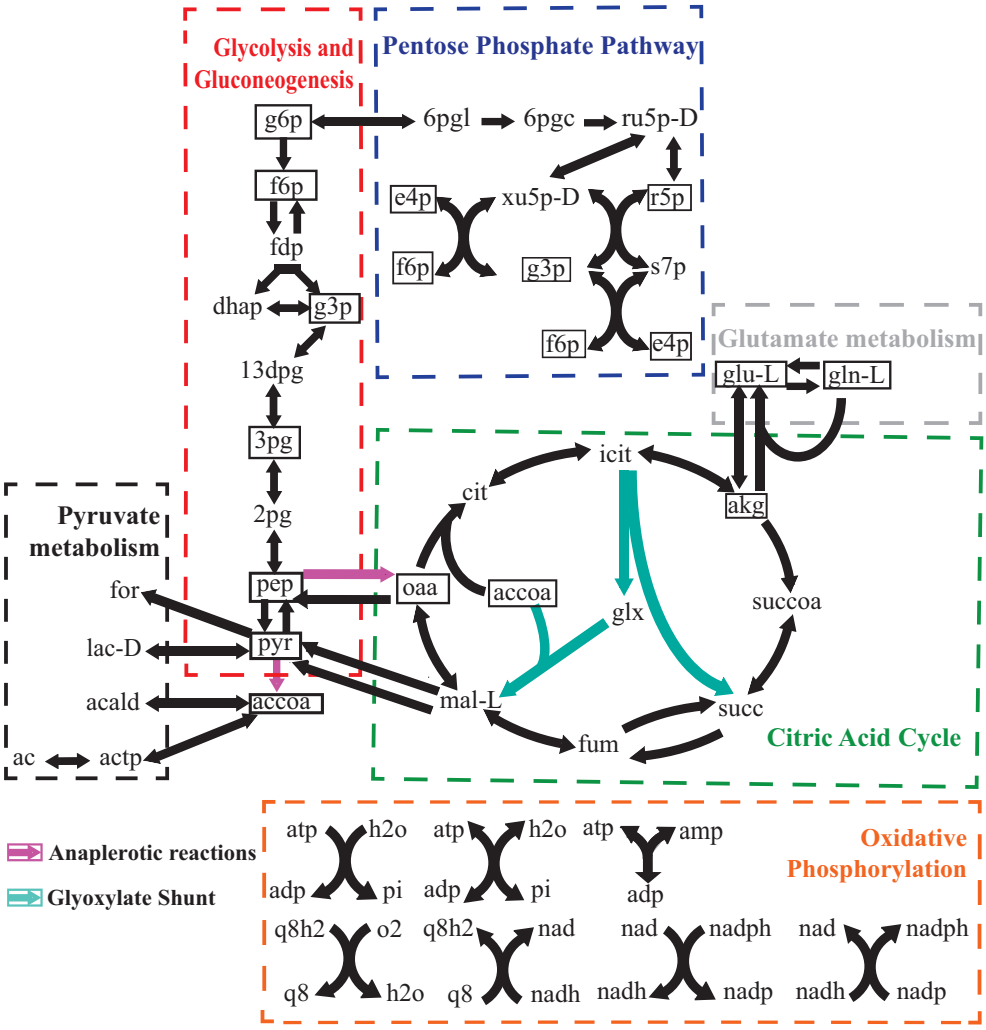
We determined the average minimal distance ($D_{min}$) between the genotype networks of phenotypes with a given complexity *(k=k'),* as a function of metabolism size *n*.

Except for the smallest metabolisms ($n<35$), the average minimal distance is close to $D_{min}$=1, regardless of phenotypic complexity. Similarly, the fraction of neighboring genotype networks is only low in small metabolisms *(n<35)*, and this fraction increases until it reaches a maximum at intermediate metabolism sizes (Figure S18). At these sizes, the fraction of neighboring genotype networks depends on the phenotypic complexity, decreasing from lowest to highest complexity. Genotype networks of phenotypes with lower phenotypic complexity tend to be closest. Figure S19 indicates that the same patterns obtain when we consider only those metabolisms where all reactions are connected to one another.

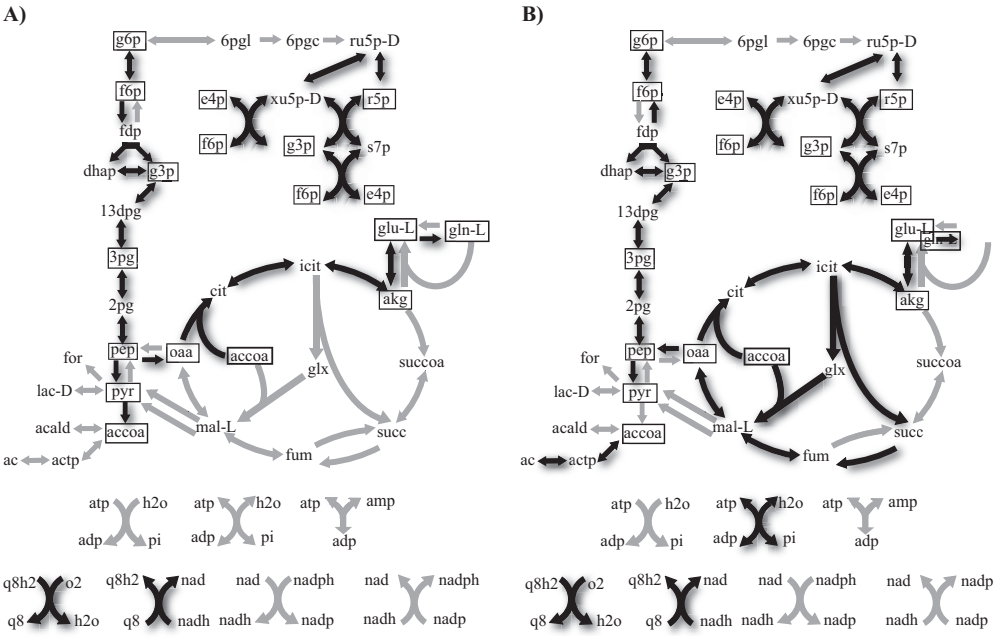
**Supplementary References:**


1. Barve A, Hosseini S-R, Martin OC, Wagner A (2014) Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. BMC Syst Biol 8: 48.
2. Hosseini S-R (2013) Exhaustive genotype-phenotype mapping in metabolic genotype space. M.Sc. Thesis, Swiss Federal Institue of Technology Zürich. Available: http://ecollection.library.ethz.ch/view/eth:7522?q=(keywords_en:PHENOTYPE. Accessed 14 November 2013.
3. Matias Rodrigues JF, Wagner A (2011) Genotype networks, innovation, and robustness in sulfur metabolism. BMC Syst Biol 5: 39.
4. Bilgin T, Wagner A (2012) Design constraints on a synthetic metabolism. PLoS One 7: e39903.

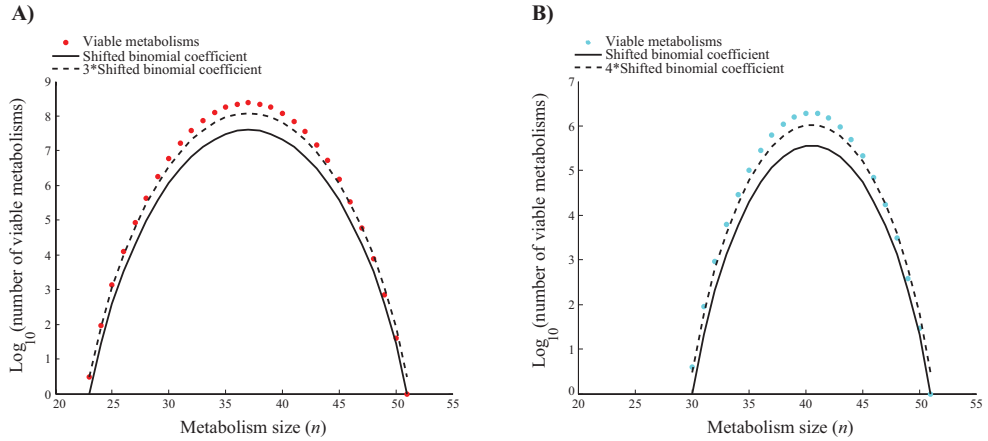
## Supplementary Figures



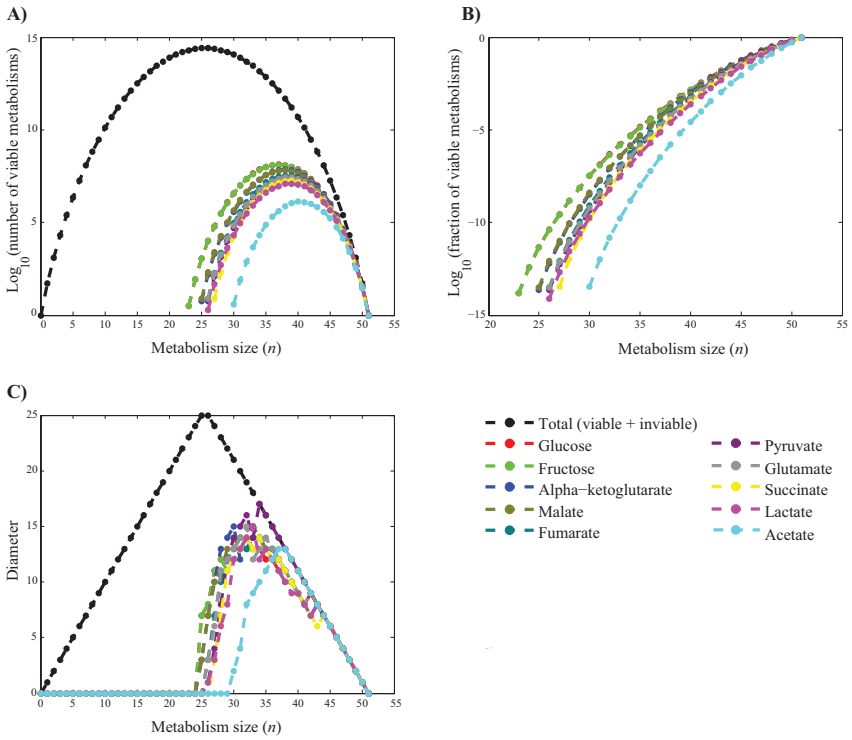

**S1 Fig. Central carbon metabolism.** Each arrow in each panel corresponds to one of the 51 internal reactions we consider. Metabolites are indicated by their acronyms (see S9 Table). Boxed metabolites correspond to 13 essential biomass precursors. Note that 4 metabolites (accoa, g3p, f6p and e4p) are shown more than once for visual clarity. Metabolic pathways, including glycolysis/gluconeogenesis, pentose-phosphate pathway, citric-acid cycle, oxidative phosphorylation, pyruvate and glutamate metabolism are distinguished by the colored and dashed rectangles. Anaplerotic reactions and glyoxylate shunt are highlighted using the purple and green arrows respectively.



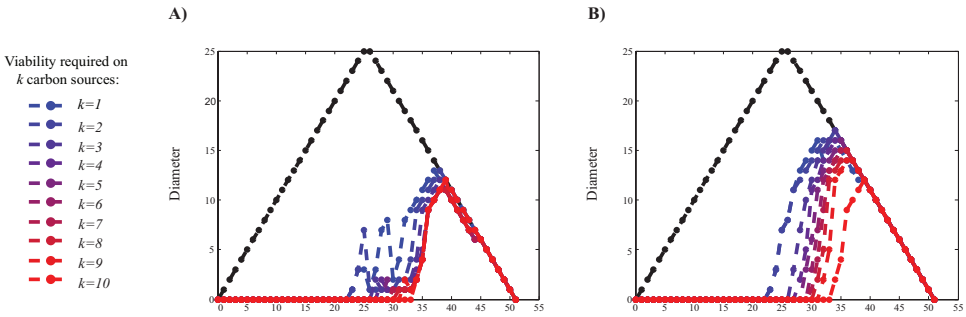

**S2 Fig. Example minimal metabolisms.** Each arrow in each panel corresponds to one of the 51 internal reactions we consider. Black arrows and gray arrows correspond to reactions that are present or absent, respectively, in the metabolism shown. Metabolites are indicated by their acronyms (see S9 Table). Boxed metabolites correspond to 13 essential biomass precursors. Note that 4 metabolites (accoa, g3p, f6p and e4p) are shown more than once for visual clarity. **(A)** Minimal metabolism with 30 reactions viable on acetate. **(B)** Minimal metabolism with 23 reactions viable on glucose.



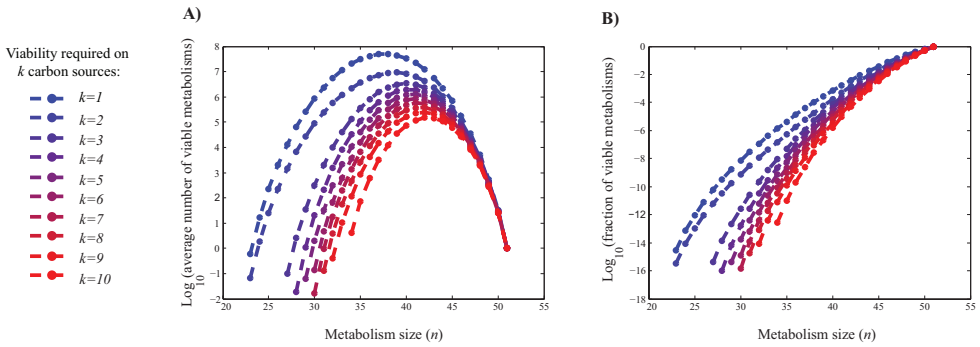

**S3 Fig. Binomial distribution of the number of viable metabolisms.** The number of metabolisms viable on **(A)** glucose (red circles) and **(B)** acetate (cyan circles) for metabolisms of a given size (x-axis). Note the logarithmic vertical axis. The black curve corresponds to the predicted number of viable metabolisms based on the shifted binomial coefficients $\binom{n-k'}{k-k'}$ where $n = 51$ and $k' = 23$ for glucose and $k' = 30$ for acetate. The dashed curve is the result of the multiplication of the black curve by a factor three in Fig **(A)** and by a factor four in Fig **(B)**.



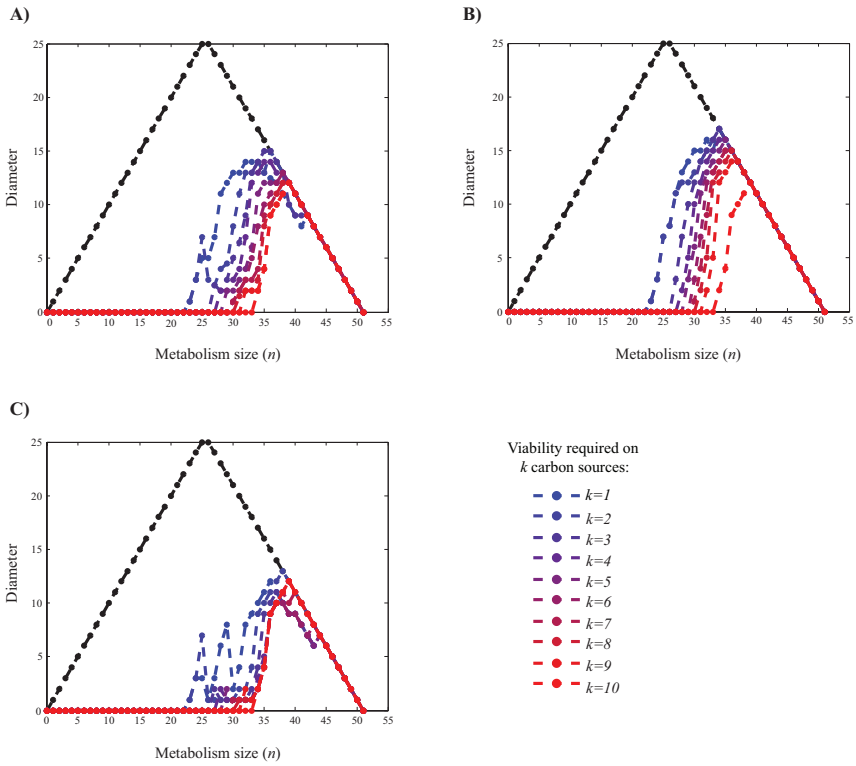

**S4 Fig. Viability of metabolisms that contain no disconnected reactions on different carbon sources and the genotypic differences among them.**
**(A)** Number of viable metabolisms lacking disconnected reactions. Black circles (vertical axis) indicate the total possible numbers of metabolisms of a given size n (horizontal axis, $N = 51$, $0 \leq n \leq 51$). Colored data points indicate the number of metabolisms without disconnected reactions that are viable on the single carbon sources indicated in the legend. **(B)** Fraction of metabolisms viable on different carbon sources. Note the logarithmic scale. **(C)** Genotype network diameter. Black circles indicate the diameter of genotype space for metabolisms of a given size, which is an upper bound for the diameter of any genotype network. Colored data points indicate genotype network for metabolisms without disconnected reactions that are viable on a single carbon source. At $n \geq 44$ all genotype networks have the maximally possible diameter.



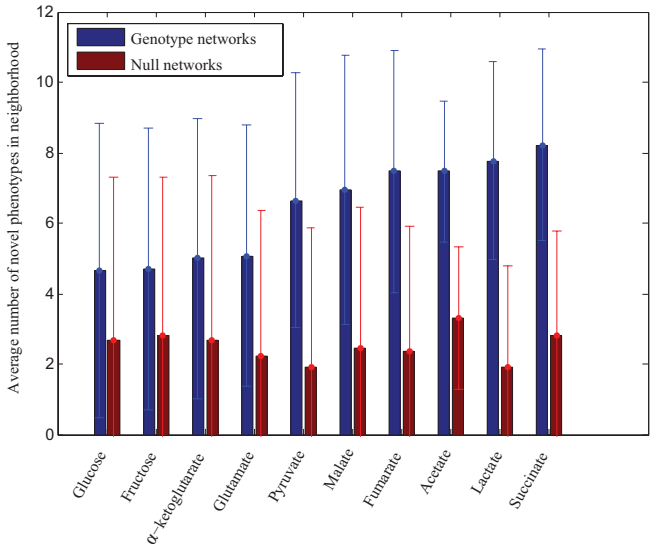

**S5 Fig. Genotype network diameter for metabolisms viable on multiple carbon sources.** Black circles indicate the diameter of genotype space for metabolisms of a given size, which is an upper bound for the diameter of any genotype network. Color-coded data points indicate in **(A)** the minimum and in **(B)** the maximum of the diameter of genotype networks for metabolisms viable on k carbon sources. At $n \geq 40$, almost all genotype networks have the maximally possible diameter.



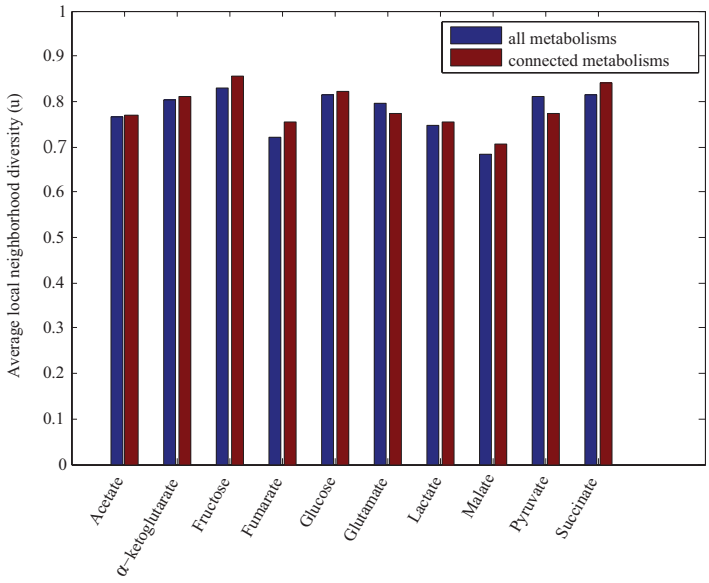

**S6 Fig. Viability of metabolisms that contain no disconnected reactions on multiple carbon sources. (A)** Average number of metabolisms that contain no disconnected reactions and are viable on k carbon sources ($1 \leq 10$, see legend). **(B)** Fraction of metabolisms without disconnected reactions that are viable on k carbon sources. Note the logarithmic vertical scale.



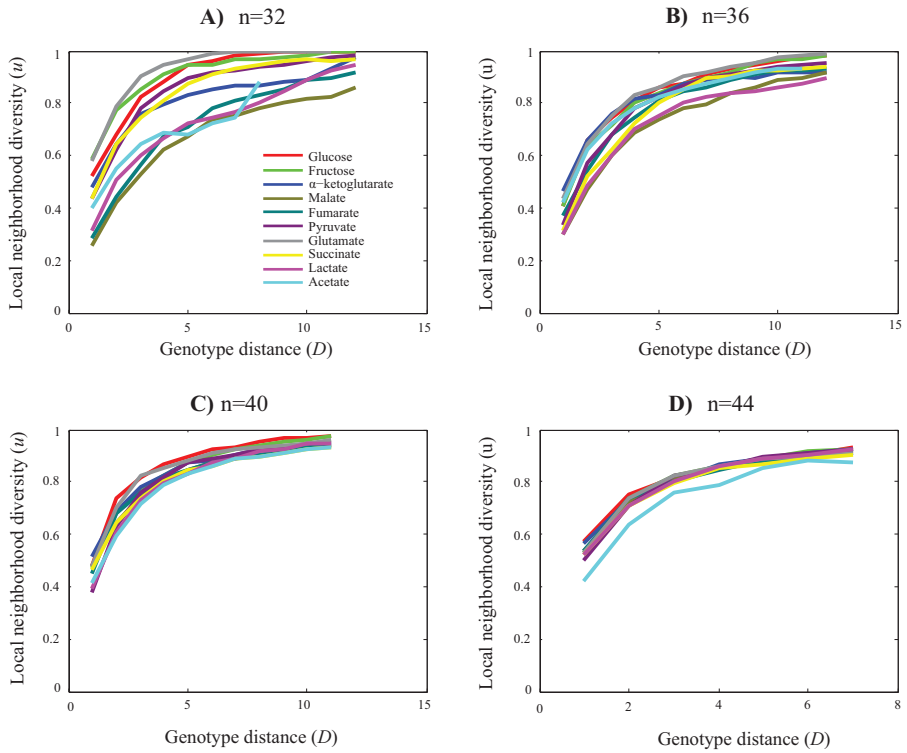

**S7 Fig. Diameter of genotype networks of metabolisms that contain no disconnected reactions and that are viable on multiple carbon sources.** Black circles indicate the diameter of genotype space for metabolisms of a given size, which is an upper bound to the diameter of any genotype network. Color-coded data points indicate **(A)** the median, **(B)** the maximum, and **(C)** the minimum of genotype network diameter for metabolisms without disconnected reactions that are viable on k carbon sources. At $n \geq 40$, almost all genotype networks have the maximally possible diameter.



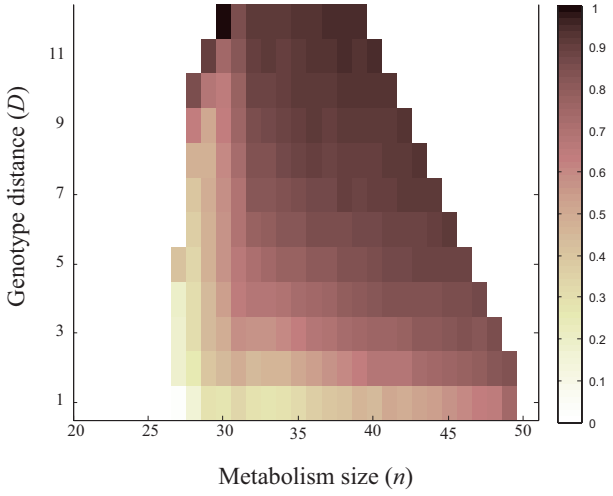

**S8 Fig. Comparison of the number of novel phenotypes in a neighborhood between actual genotype networks and randomized (null) networks.** We generated null networks for each genotype network by randomly permuting the phenotypes among all members of the genotype network, and did so for all genotype networks. The height of each bar (blue: actual genotype network; red: randomized network) corresponds to the average number of novel phenotypes in a neighborhood, where the average is taken over all metabolisms viable on the carbon source indicated along the horizontal axis. Error bars indicate one standard deviation.



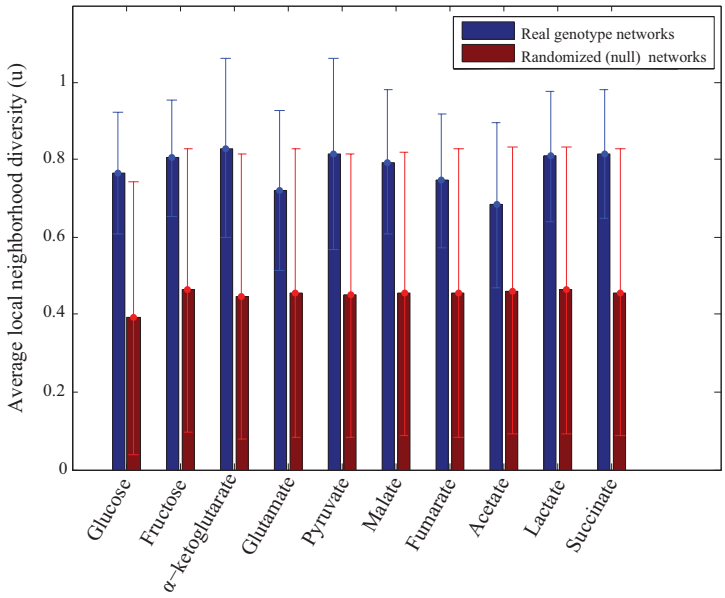

**S9 Fig. Average local neighborhood diversity.** The height of the bars corresponds to the average local neighborhood diversity (u) among all pairs of metabolisms viable on the carbon source indicated along the horizontal axis. Blue bars are based on all viable metabolisms, and brown bars are based on viable metabolisms lacking any disconnected reactions.



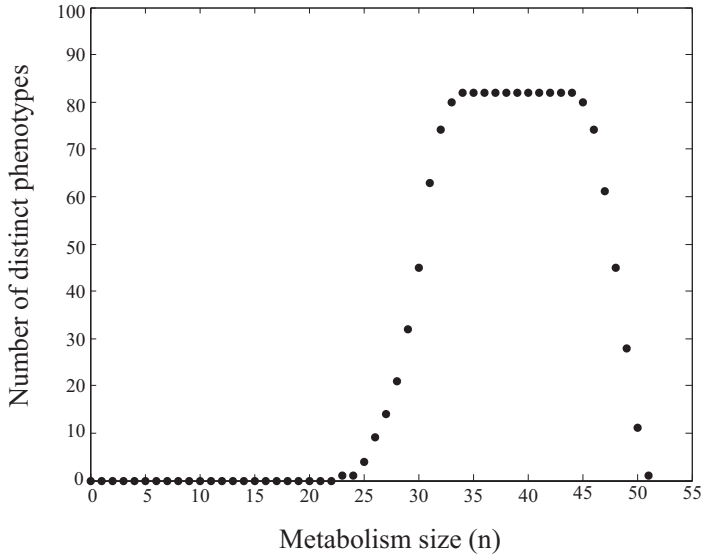

**S10 Fig. Local neighbor diversity as a function of genotypic distance.** Local neighborhood diversity (u) as a function of genotypic distance (D) is shown for metabolisms viable on a given carbon source (color-coded, see legend). **(A)** Data for metabolisms with n = 32, **(B)** n = 36, **(C)** n = 40, and **(D)** n = 44 reactions.



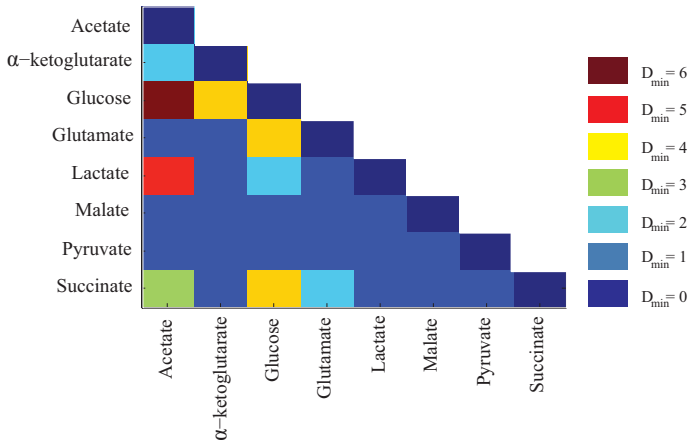

**S11 Fig. Average local neighborhood diversity.** For metabolism pairs of a given genotype distance (D, y-axis) and size (n, x-axis), that are viable on fumarate, the average local neighborhood diversity (u) (see color legend) is shown. Data are based on 1000 randomly sampled networks for each metabolism size (n), and genotypic distance (D).



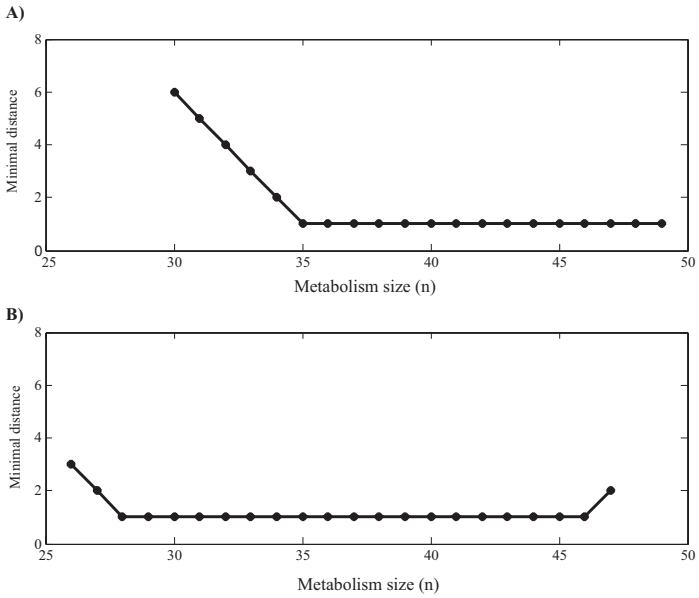

**S12 Fig. Comparison of the local neighborhood diversity (u) between actual genotype networks and randomized (null) networks.**
We generated null networks for each genotype network by randomly permuting the phenotypes among all members of the genotype network, and did so for all genotype networks. The height of each bar (blue: actual genotype network; brown: randomized network) corresponds to the average local neighborhood diversity (u), where the average is taken over all metabolisms viable on the carbon source indicated along the horizontal axis. Error bars indicate one standard deviation.



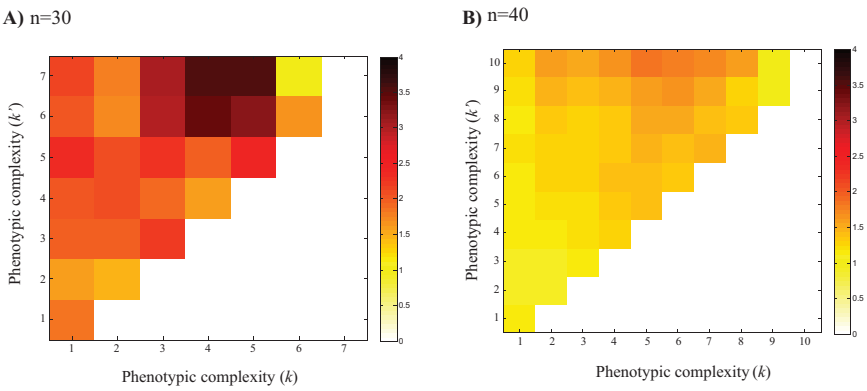

**S13 Fig. Number of phenotypes as a function of metabolism size (n).** Each black circle shows the number of phenotypes (out of $2^{10} = 1024$ possible phenotypes) for which at least one metabolism with this phenotype exists (vertical axis), as a function of the size n of the metabolism (horizontal axis).



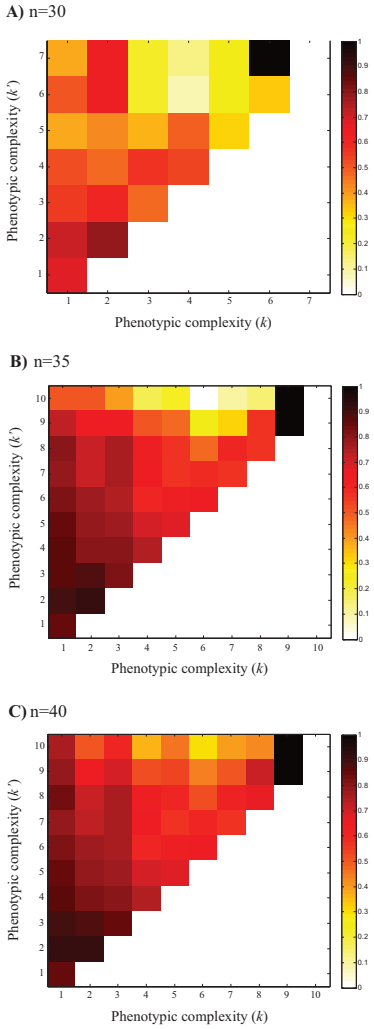

**S14 Fig. Minimal distance between genotype networks.** Each rectangular colored area shows the color-coded minimal genotype distance $D_{min}$ between genotype networks for phenotypes viable on carbon sources indicated on the corresponding $(x, y)$ positions ($n = 30$).



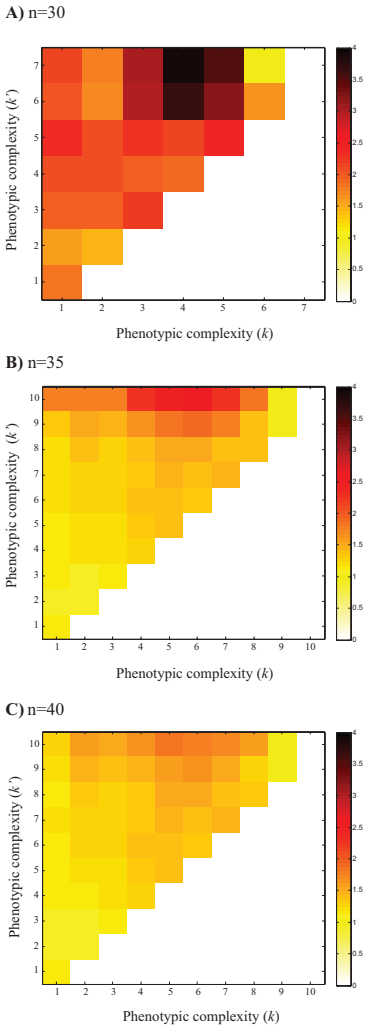

**S15 Fig. Minimal distance between genotype networks as a function of metabolism size.** Minimal distance between pairs of genotype networks for metabolisms **(A)** viable only on acetate or glucose, and **(B)** viable only on glutamate or malate, as a function of metabolism size (horizontal axes). Note that at the largest n where metabolisms exist that are viable on only a single carbon source, $D_{min}$ increases in some cases from one to two and that is because of the small size of genotype networks at the largest metabolism sizes.



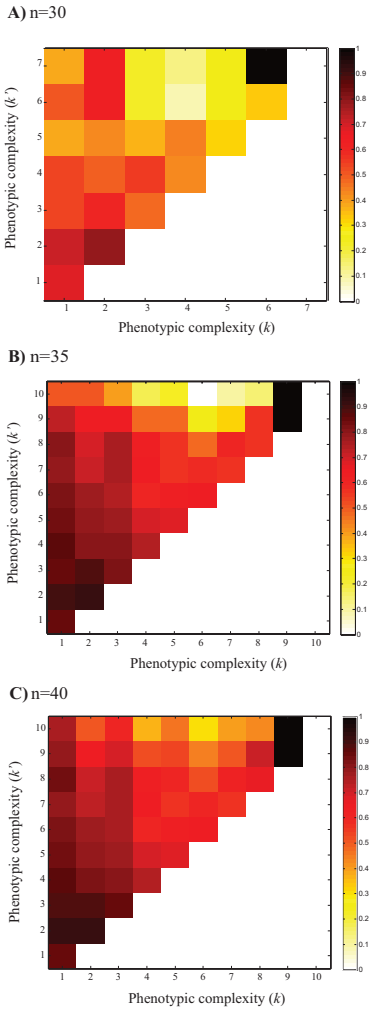

**S16 Fig. Minimal distance between genotype networks as a function of phenotypic complexity.** Each rectangular area shows the color-coded average of the minimal distance $D_{min}$ (see color legend) between pairs of genotype networks with phenotypic complexity (k, x-axis) and (k', y-axis) for metabolisms of size 34 **(A)** $n = 30$ and **(B)** $n = 40$. At $n = 30$, no metabolism is viable on more than $k = k' = 7$ carbon sources, hence no distances can be calculated for the corresponding values of k. The highest average minimal distances exists for metabolisms of complexity **(A)** $(k, k') = (4, 7)$, which show $D_{min} = 3.5$ and **(B)** (k, k')=(10, 7), which show $D_{min} = 1.82$.



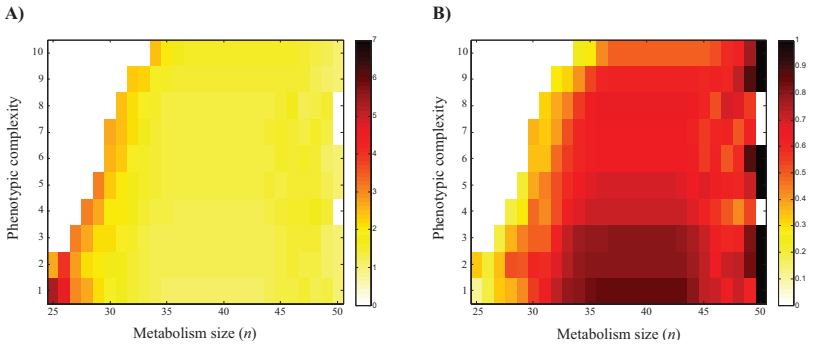

**S17 Fig. Fraction of neighboring genotype networks as a function of phenotypic complexity.** Each color-coded rectangular area shows the fraction of genotype networks that are neighbors (have $D_{min} = 1$) among pairs of genotype networks of a given phenotypic complexity (k, x-axis) and (k', yaxis), for metabolisms of size **(A)** $n = 30$ **(B)** $n = 35$ and **(C)** $n = 40$. At $n = 30$, no metabolism is viable on more than $k = k' = 7$ carbon sources, hence no distances can be calculated for the corresponding values of k.
(EPS)



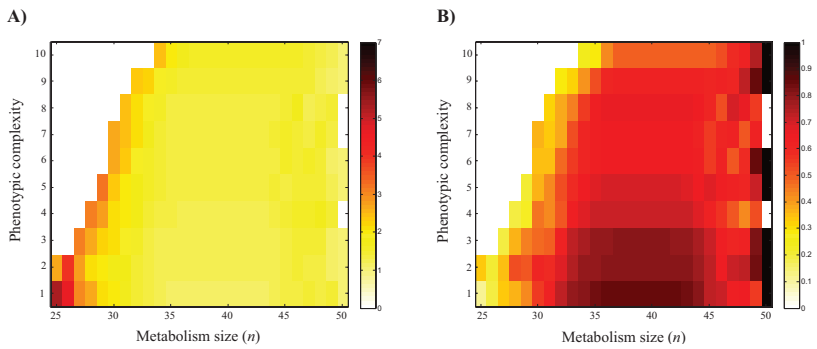

**S18 Fig. Minimal distance between genotype networks as a function of phenotypic complexity for metabolisms without disconnected reactions.** Each color-coded rectangular area shows the average minimal distance of pairs of genotype networks with phenotypic complexity (k, x-axis) and (k', y-axis), for metabolisms of size **(A)** $n = 30$ **(B)** $n = 35$ and **(C)** $n = 40$. Note that in this analysis (unlike Figs 5C and S14) only genotype networks of metabolisms without disconnected reactions are considered. At $n = 30$, no metabolism is viable on more than $k = k' = 7$ carbon sources, hence no distances can be calculated for the corresponding values of k. The highest average minimal distances exists for metabolisms of complexity **(A)** $(k, k') = (4, 7)$, which show $D_{min} = 3.87$, **(B)** $(k, k') = (10, 6)$, which show $D_{min} = 2.6$ and **(C)** $(k, k') = (10, 5)$, which show $D_{min} = 1.82$.



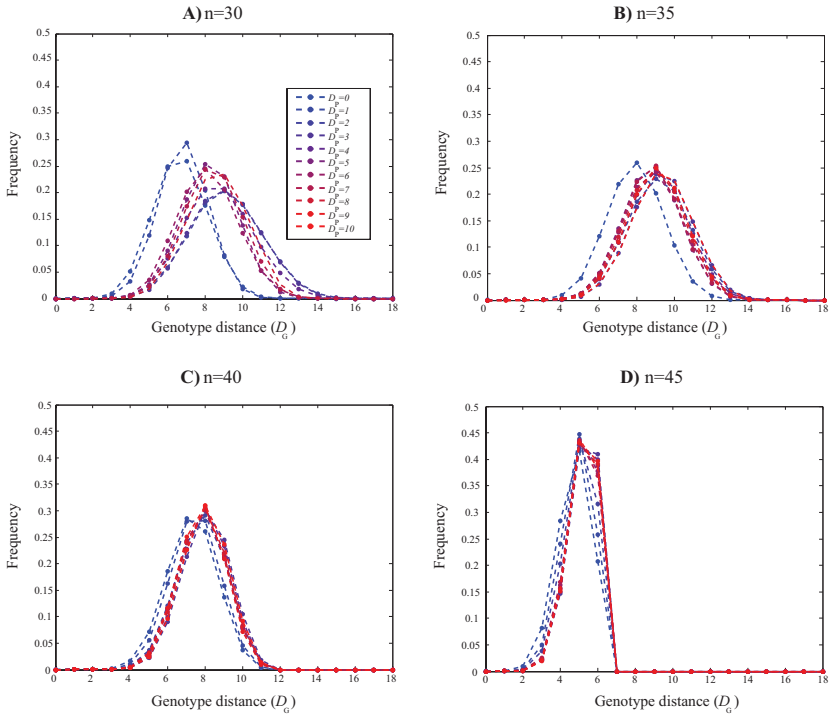

**S19 Fig. Fraction of neighboring genotype networks as a function of phenotypic complexity.** Each color-coded rectangular area shows the fraction of neighboring genotype networks among pairs of genotype networks with a given phenotypic complexity (k, x-axis) and (k', y-axis), for metabolisms of size **(A)** n=30 **(B)** n=35 and **(C)** n=40. Note that in this analysis (unlike S15 Fig) only genotype networks of metabolisms without disconnected reactions are considered. At $n = 30$, no metabolism is viable on more than $k = k' = 7$ carbon sources, hence no distances can be calculated for the corresponding values of k.



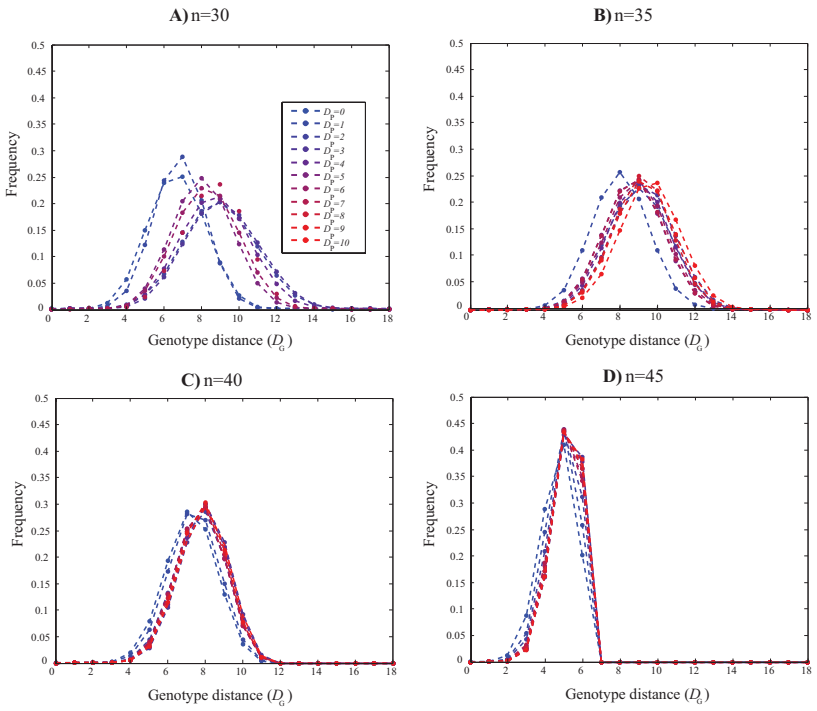

**S20 Fig. Average minimal distance and fraction of neighboring genotype networks as a function of metabolism size.** Each color-coded rectangular area indicates **(A)** the average minimal distance, and **(B)** the average fraction of neighboring genotype networks, for all pairs of genotype networks of metabolisms with a given size (n, x-axis) and with a given phenotypic complexity (y-axis).



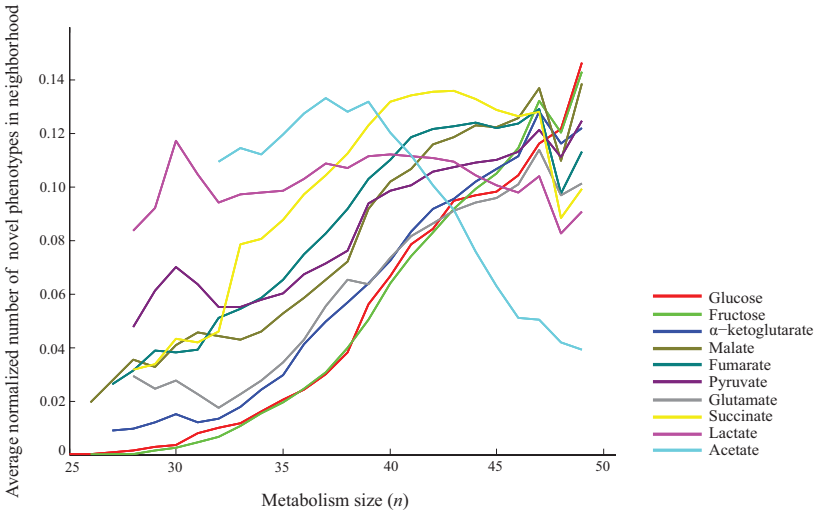

**S21 Fig. Average minimal distance and fraction of neighboring genotype networks as a function of metabolism size.** Each color-coded rectangular area indicates **(A)** the average minimal distance, and **(B)** the average fraction of neighboring genotype networks, for all pairs of genotype networks of metabolisms with a given size (n, x-axis) and with a given phenotypic complexity (y-axis). Note that in this analysis (unlike S18 Fig) only genotype networks of metabolisms without disconnected reactions are considered.




**S22 Fig. Distribution of genotypic distances $D_G$ for phenotypes of a given distance $D_P$.** Data points of given shading indicate the distribution of genotypic distance $D_G$ (x-axis) among pairs of metabolisms with a given phenotypic distance $D_P$, as indicated in the color legend for metabolisms of size **(A)** $n = 30$, **(B)** $n = 35$, **(C)** $n = 40$ and **(D)** $n = 45$.




**S23 Fig. Distribution of genotypic distances $D_G$ for phenotypes of a given distance $D_P$ (only connected metabolisms).** Data points of a given shading indicate the distribution of genotypic distance $D_G$ (x-axis) among pairs of metabolisms with a given phenotypic distance $D_P$, as indicated in the color for metabolisms of size **(A)** n = 30, **(B)** n = 35, **(C)** n = 40 and **(D)** n = 45. Note that in this analysis (unlike S20 Fig.) only genotype networks of metabolisms without disconnected reactions are considered.




**S24 Fig. Normalized number of novel phenotypes in neighborhood.** Average number of distinct novel metabolic phenotypes in neighborhood of metabolisms of a given size (n) and viable on a given carbon source, divided by the number of distinct phenotypes existing in metabolisms of this size.

| Carbon source | $n_{min}$ | number of viable metabolisms at $n_{min}$ | total number of viable metabolisms | $n_{max}$ | maximum number of viable metabolisms at $n_{max}$ |
|---|---|---|---|---|---|
| Acetate | 30 | 4 | 10850304 | 40 | 1922772 |
| α-ketoglutarate | 25 | 6 | 928636928 | 38 | 153139863 |
| Fructose | 23 | 3 | 1473094400 | 37 | 227234491 |
| Fumarate | 26 | 8 | 412139520 | 38 | 68804412 |
| Glucose | 23 | 3 | 1549771520 | 37 | 239328665 |
| Glutamate | 26 | 8 | 342265856 | 38 | 56385611 |
| Lactate | 26 | 2 | 141944832 | 39 | 23190620 |
| Malate | 25 | 8 | 781627392 | 38 | 126729791 |
| Pyruvate | 26 | 6 | 353501184 | 38 | 57805322 |
| Succinate | 27 | 8 | 217774080 | 39 | 36626028 |

**S1 Table. Further information on the number of viable metabolisms and minimal metabolisms.** Each row contains data for metabolisms viable on one of 10 carbon sources, as indicated in the left-most column. Columns from left to right indicate the minimum number of reactions required to be viable on that carbon source ($n_{min}$), the number of metabolisms with this minimum number of reactions, the total number of viable metabolisms, the size $n_{max}$ at which the number of viable metabolisms is maximal, and the maximum number of viable metabolisms at $n_{max}$.

| $n$ | Acetate | | Alpha-ketoglutarate | | Fructose | | Fumarate | | Glucose | | Glutamate | | Lactate | | Malate | | Pyruvate | | Succinate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ |
| 23 | | | | | 2 | 0.6666 | | | 2 | 0.6666 | | | | | | | | | | |
| 24 | | | | | 2 | 0.6373 | | | 2 | 0.6373 | | | | | | | | | | |
| 25 | | | 1 | 1 | 2 | 0.9969 | | | 2 | 0.9969 | | | | | 1 | 1 | | | | |
| 26 | | | 2 | 0.9672 | 2 | 0.9919 | 1 | 1 | 2 | 0.9920 | 1 | 1 | 2 | 1 | 2 | 0.9642 | 3 | 0.3333 | | |
| 27 | | | 2 | 0.9973 | 1 | 1 | 2 | 0.9629 | 1 | 1 | 3 | 0.9425 | 1 | 0.9622 | 1 | 1 | 4 | 0.9617 | 1 | 1 |
| 28 | | | 3 | 0.9996 | 2 | 0.9999 | 1 | 1 | 2 | 0.9999 | 3 | 0.9977 | 2 | 1 | 1 | 1 | 3 | 0.9964 | 2 | 0.9649 |
| 29 | | | 2 | 0.9999 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9999 | 3 | 0.9992 | 1 | 1 | 3 | 0.9974 | 1 | 1 |
| 30 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9998 | 1 | 1 | 3 | 0.9999 | 1 | 1 |
| 31 | 2 | 0.9545 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9999 | 1 | 1 | 1 | 1 | 1 | 1 |

**S2 Table. Number of connected components and the fractional size of the largest component for the metabolisms viable on a given carbon source.** Each pair of columns shows the number of connected components ($n_C$) as well as the fractional size of the largest component ($r_G$) for metabolism of different sizes n (left-most column) that are required to be viable on a given carbon source (first row). For metabolisms with $n \geq 32$, for every carbon source, $n_C$ and $r_G$ both equal to one. Empty fields mean that no viable metabolism exists for the corresponding carbon source and metabolism size n.

| Metabolism size | Acetate | Alphaketoglutarate | Fructose | Fumarate | Glucose | Glutamate | Lactate | Malate | Pyruvate | Succinate |
|---|---|---|---|---|---|---|---|---|---|---|
| 23 | | | 100.0000 | | 100.0000 | | | | | |
| 24 | | | 96.0000 | | 96.0000 | | | | | |
| 25 | | 100.0000 | 100.0000 | | 100.0000 | | | 100.0000 | | |
| 26 | | 50.0000 | 96.5600 | 100.0000 | 96.1600 | 75.0000 | 100.0000 | 100.0000 | 100.0000 | |
| 27 | | 37.8378 | 100.0000 | 100.0000 | 100.0000 | 60.7143 | 100.0000 | 88.2353 | 90.8696 | 100.0000 |
| 28 | | 29.1262 | 96.6543 | 87.7551 | 98.0000 | 54.2636 | 100.0000 | 70.5871 | 73.3333 | 50.0000 |
| 29 | | 19.9496 | 71.9000 | 50.0000 | 49.7200 | 50.0700 | 99.1667 | 48.3871 | 95.4071 | 75.0000 |
| 30 | 100.0000 | 22.9630 | 68.1300 | 46.3612 | 46.8557 | 42.3800 | 94.1176 | 48.2759 | 66.6667 | 54.7896 |
| 31 | 88.8889 | 04.2300 | 61.0700 | 34.9593 | 40.6250 | 33.3700 | 86.2700 | 37.5465 | 49.2063 | 42.6200 |
| 32 | 84.6154 | 11.2100 | 62.2174 | 25.4144 | 56.4103 | 02.8798 | 68.9446 | 48.6842 | 50.0000 | 35.2000 |
| 33 | 65.1556 | 22.2341 | 58.0000 | 54.1604 | 55.6818 | 10.6900 | 71.5284 | 65.1163 | 56.1538 | 53.2000 |
| 34 | 49.7000 | 44.5946 | 20.0000 | 51.0204 | 58.2800 | 26.8536 | 72.9299 | 90.0000 | 57.6923 | 51.8047 |
| 35 | 36.4300 | 61.8182 | 58.8837 | 59.2857 | 59.7800 | 48.2759 | 52.1739 | 54.6032 | 53.4483 | 30.7692 |
| 36 | 26.6200 | 58.7124 | 61.6600 | 57.5800 | 61.2300 | 63.5406 | 53.1469 | 58.8900 | 60.6626 | 53.1282 |
| 37 | 16.4163 | 65.7900 | 67.4800 | 61.0400 | 65.2400 | 58.5200 | 55.4600 | 61.9700 | 62.8800 | 58.9300 |
| 38 | 80.0000 | 70.2400 | 69.5800 | 66.4600 | 70.6000 | 64.5500 | 60.3300 | 68.0000 | 66.8000 | 64.5100 |
| 39 | 47.9769 | 76.0176 | 49.2200 | 42.3000 | 48.4400 | 45.5100 | 40.9200 | 45.3900 | 45.5400 | 41.3400 |
| 40 | 30.6100 | 57.6500 | 54.8100 | 49.9900 | 55.6100 | 52.9600 | 46.9700 | 52.9700 | 52.5100 | 48.7000 |
| 41 | 39.1100 | 63.6400 | 60.8800 | 59.2200 | 61.3500 | 60.7300 | 54.3000 | 60.3300 | 59.1100 | 56.0800 |
| 42 | 48.5400 | 70.2800 | 67.7500 | 65.9600 | 69.5300 | 67.8000 | 60.3100 | 67.5100 | 65.3300 | 64.0900 |
| 43 | 58.3500 | 75.9000 | 75.7300 | 73.4600 | 76.1000 | 73.5900 | 68.5000 | 74.7200 | 71.9200 | 70.7400 |
| 44 | 67.7000 | 81.5700 | 81.3900 | 79.1900 | 82.4000 | 80.4400 | 75.2100 | 81.4800 | 78.7900 | 78.2900 |
| 45 | 76.6500 | 86.7100 | 87.1000 | 85.6200 | 86.8400 | 84.9200 | 82.9000 | 86.0500 | 85.1000 | 84.2800 |
| 46 | 84.2100 | 93.2300 | 93.2600 | 92.3100 | 91.7900 | 92.7100 | 88.2800 | 92.6300 | 92.8500 | 88.9400 |
| 47 | 90.0700 | 94.1000 | 94.5700 | 93.5600 | 95.0100 | 93.9500 | 93.0300 | 94.5200 | 93.9100 | 93.7000 |
| 48 | 94.4000 | 96.8400 | 97.0200 | 96.5400 | 97.0900 | 96.6400 | 96.3900 | 97.0700 | 97.1500 | 96.2500 |
| 49 | 97.8800 | 98.7600 | 98.9900 | 98.7000 | 98.9000 | 98.7100 | 98.5600 | 98.8500 | 98.6800 | 98.3500 |
| 50 | 99.5074 | 99.7155 | 99.7301 | 99.6825 | 99.7436 | 99.6825 | 99.6639 | 99.6997 | 99.7155 | 99.6639 |

**S3 Table. Percentage of metabolism pairs with genotypic distance equal to network diameter, where different members of a metabolism pair do not use alternative reactions that differ only in a co-factor.** Each entry of the table corresponds to the genotype network of metabolisms of a given size (specified by the left-most column) that are viable on a given carbon source (specified in the first row). Empty fields mean that no viable metabolism exists for the corresponding carbon source and metabolism size.

| Metabolism size | Acetate | Alphaketoglutarate | Fructose | Fumarate | Glucose | Glutamate | Lactate | Malate | Pyruvate | Succinate |
|---|---|---|---|---|---|---|---|---|---|---|
| 23 | | | 0 | | 0 | | | | | |
| 24 | | | 0 | | 0 | | | | | |
| 25 | | 0 | 0 | | 0 | | | 0 | | |
| 26 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 27 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 28 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 29 | | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 30 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 31 | 0 | 1 | 1 | 0 | 2 | 0 | 0 | 2 | 2 | 0 |
| 32 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 2 | 0 |
| 33 | 0 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| 34 | 0 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 2 | 1 |
| 35 | 0 | 2 | 2 | 1 | 1 | 0 | 2 | 2 | 2 | 2 |
| 36 | 0 | 1 | 2 | 1 | 2 | 0 | 2 | 1 | 2 | 1 |
| 37 | 0 | 1 | 1 | 1 | 2 | 0 | 2 | 1 | 2 | 1 |
| 38 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 |
| 39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 42 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 43 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 49 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 50 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**S4 Table. Minimum number of blocked reactions among all pairs of metabolisms with genotypic distance equal to genotype network diameter.** Each element of the table corresponds to the genotype network of metabolisms of a given size (specified in the left-most column) that are viable on a given source (specified in the first row). Empty fields mean that no viable metabolism exists for the corresponding carbon source and metabolism size n.

| $n$ | Acetate | | Alpha-ketoglutarate | | Fructose | | Fumarate | | Glucose | | Glutamate | | Lactate | | Malate | | Pyruvate | | Succinate | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ |
| 23 | | | | | 2 | 0.6666 | | | 2 | 0.6666 | | | | | | | | | | |
| 24 | | | | | 2 | 0.6190 | | | 2 | 0.6190 | | | | | | | | | | |
| 25 | | | | 1 | 2 | 0.9964 | | | 2 | 0.9965 | | | | | 1 | 1 | | | | |
| 26 | | | 2 | 0.9622 | 2 | 0.9907 | 1 | 1 | 2 | 0.9908 | 1 | 1 | 2 | 1 | 2 | 0.96 | 3 | 0.3333 | | |
| 27 | | | 2 | 0.9833 | 1 | 1 | 2 | 0.9583 | 1 | 1 | 3 | 0.9322 | 1 | 0.9574 | 1 | 0.9983 | 4 | 0.6564 | 1 | 1 |
| 28 | | | 3 | 0.9994 | 2 | 0.9999 | 1 | 0.9982 | 2 | 0.9999 | 3 | 0.9869 | 2 | 0.9962 | 1 | 1 | 3 | 0.7072 | 2 | 0.9607 |
| 29 | | | 2 | 0.9999 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9998 | 3 | 0.9989 | 1 | 1 | 3 | 0.7352 | 1 | 0.9983 |
| 30 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9989 | 1 | 1 | 3 | 0.7698 | 1 | 1 |
| 31 | 2 | 0.95 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9998 | 1 | 1 | 1 | 0.8041 | 1 | 1 |
| 32 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.8386 | 1 | 1 |
| 33 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.8717 | 1 | 1 |
| 34 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9016 | 1 | 1 |
| 35 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9273 | 1 | 1 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9480 | 1 | 1 |
| 37 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9640 | 1 | 1 |
| 38 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9758 | 1 | 1 |
| 39 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9842 | 1 | 1 |
| 40 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9901 | 1 | 1 |
| 41 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9940 | 1 | 1 |
| 42 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9965 | 1 | 1 |
| 43 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9981 | 1 | 1 |
| 44 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9990 | 1 | 1 |
| 45 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9996 | 1 | 1 |
| 46 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9998 | 1 | 1 |
| 47 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9999 | 1 | 1 |

**S5 Table. Number of connected components and the fractional size of the largest component for the metabolisms viable on a given carbon source.** Each pair of columns shows the number of connected components ($n_C$) as well as the fractional size of the largest component ($r_G$) for metabolisms of different size n (left-most column) that are viable on a given carbon source (first row). For metabolisms with $n \geq 48$ for every carbon source, $n_C$ and $r_G$ both equal to one. Empty fields mean that no viable metabolism exists for the corresponding carbon source and n. The information in this table is restricted to metabolisms without disconnected reactions.

| | Number *k* of carbon sources | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number *n* of reactions | *k = 1* | | *k = 2* | | *k = 3* | | *k = 4* | | *k = 5* | | *k = 6* | | *k = 7* | | *k = 8* | | *k = 9* | | *k = 10* | |
| | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ |
| 23 | 2 | 0.6666 | 2 | 0.6666 | | | | | | | | | | | | | | | | |
| 24 | 2 | 0.6373 | 2 | 0.6373 | | | | | | | | | | | | | | | | |
| 25 | 2 | 0.9969 | 2 | 0.9969 | | | | | | | | | | | | | | | | |
| 26 | 2 | 0.9672 | 1 | 1 | | | | | | | | | | | | | | | | |
| 27 | 2 | 0.9629 | 1 | 1 | 1 | 1 | | | | | | | | | | | | | | |
| 28 | 2 | 0.9996 | 2 | 0.9606 | 2 | 0.9529 | | | | | | | | | | | | | | |
| 29 | 2 | 0.9995 | 2 | 0.9994 | 2 | 0.9593 | 1 | 1 | 1 | 1 | | | | | | | | | | |
| 30 | 3 | 0.9999 | 2 | 0.9929 | 2 | 0.9591 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | | | | | | |
| 31 | 2 | 0.9772 | 2 | 0.9993 | 2 | 0.9905 | 2 | 0.9221 | 2 | 0.9937 | 1 | 1 | 1 | 1 | 1 | 1 | | | | |
| 32 | 1 | 1 | 2 | 0.9999 | 2 | 0.9979 | 3 | 0.9851 | 3 | 0.9811 | 2 | 0.9851 | 1 | 1 | 1 | 1 | 1 | 1 | | |
| 33 | 1 | 1 | 1 | 1 | 2 | 0.9998 | 3 | 0.9973 | 3 | 0.9925 | 3 | 0.991 | 2 | 0.9984 | 1 | 1 | 1 | 1 | | |
| 34 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9998 | 3 | 0.9976 | 3 | 0.9966 | 3 | 0.9862 | 3 | 0.9966 | 1 | 1 | 1 | 1 |
| 35 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9995 | 2 | 0.9973 | 3 | 0.9963 | 3 | 0.9113 | 3 | 0.9113 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0.9959 | 2 | 0.9959 |

**S6 Table. Median of the number of connected components and the fractional size of the largest component for metabolisms viable on multiple carbon sources.** Each pair of columns shows the median number of connected components ($n_C$), as well as the fractional size of the largest component ($r_G$) for metabolisms of different sizes n (left-most column) that are required to be viable on k carbon sources (first row). For metabolisms with $n \geq 37$, for every k, $n_C$ and $r_G$ both equal to one. Empty fields mean that no viable metabolism exists for the corresponding k and n.

| | Number $k$ of carbon sources | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number $n$ of reactions | $k = 1$ | | $k = 2$ | | $k = 3$ | | $k = 4$ | | $k = 5$ | | $k = 6$ | | $k = 7$ | | $k = 8$ | | $k = 9$ | | $k = 10$ | |
| | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ |
| 23 | 2 | 0.6666 | 2 | 0.6666 | | | | | | | | | | | | | | | | |
| 24 | 2 | 0.6373 | 2 | 0.6373 | | | | | | | | | | | | | | | | |
| 25 | 2 | 0.9969 | 2 | 0.9969 | | | | | | | | | | | | | | | | |
| 26 | 3 | 0.992 | 2 | 1 | | | | | | | | | | | | | | | | |
| 27 | 4 | 1 | 2 | 1 | 1 | 1 | | | | | | | | | | | | | | |
| 28 | 3 | 1 | 4 | 1 | 4 | 1 | 1 | 1 | | | | | | | | | | | | |
| 29 | 3 | 0.9999 | 4 | 1 | 4 | 1 | 4 | 1 | 1 | 1 | | | | | | | | | | |
| 30 | 3 | 0.9999 | 5 | 1 | 5 | 1 | 5 | 1 | 5 | 1 | 1 | 1 | 1 | 1 | | | | | | |
| 31 | 2 | 0.9999 | 5 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 3 | 1 | 1 | 1 | | | | |
| 32 | 1 | 1 | 6 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 4 | 1 | 1 | 1 | | |
| 33 | 1 | 1 | 3 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 7 | 1 | 6 | 1 | 5 | 1 | 4 | 1 | | |
| 34 | 1 | 1 | 3 | 1 | 5 | 1 | 7 | 1 | 7 | 1 | 6 | 1 | 6 | 1 | 6 | 1 | 5 | 1 | 1 | 1 |
| 35 | 1 | 1 | 1 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 1 | 3 | 0.9995 | 3 | 0.9113 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 | 0.9959 |

**S7 Table. Maximum number of connected components and the fractional size of the largest component for metabolisms viable on multiple carbon sources.** Each pair of columns shows the maximum number of connected components ($n_C$) as well as the fractional size of the largest component ($r_G$) for metabolism of different sizes n (left-most column) and required to be viable on k carbon sources (first row). For metabolisms with $n \geq 37$, for every k, $n_C$ and $r_G$ both equal to one. Empty fields mean that no viable metabolism exists for the corresponding k and n.

| Number $n$ of reactions | $k = 1$ | | $k = 2$ | | $k = 3$ | | $k = 4$ | | $k = 5$ | | $k = 6$ | | $k = 7$ | | $k = 8$ | | $k = 9$ | | $k = 10$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ | $n_C$ | $r_G$ |
| 23 | 2 | 0.6666 | 2 | 0.6666 | | | | | | | | | | | | | | | | |
| 24 | 2 | 0.6373 | 2 | 0.6373 | | | | | | | | | | | | | | | | |
| 25 | 2 | 0.9969 | 2 | 0.9969 | | | | | | | | | | | | | | | | |
| 26 | 2 | 0.3333 | 1 | 0.9919 | | | | | | | | | | | | | | | | |
| 27 | 1 | 0.9425 | 1 | 0.5 | 1 | 1 | | | | | | | | | | | | | | |
| 28 | 1 | 0.9649 | 1 | 0.2857 | 1 | 0.2857 | 1 | 1 | | | | | | | | | | | | |
| 29 | 2 | 0.9974 | 1 | 0.5049 | 1 | 0.2857 | 1 | 0.2857 | 1 | 1 | | | | | | | | | | |
| 30 | 3 | 0.9998 | 1 | 0.7142 | 1 | 0.3809 | 1 | 0.3809 | 1 | 0.3809 | 1 | 1 | 1 | 1 | | | | | | |
| 31 | 2 | 0.9545 | 1 | 0.7377 | 1 | 0.4 | 1 | 0.3809 | 1 | 0.3809 | 1 | 0.3809 | 1 | 0.6666 | 1 | 1 | | | | |
| 32 | 1 | 1 | 1 | 0.9518 | 1 | 0.4122 | 1 | 0.4122 | 1 | 0.4122 | 1 | 0.4122 | 1 | 0.6666 | 1 | 0.8256 | 1 | 1 | | |
| 33 | 1 | 1 | 1 | 0.9268 | 1 | 0.9232 | 1 | 0.6043 | 1 | 0.6043 | 1 | 0.6043 | 1 | 0.6043 | 1 | 0.8076 | 1 | 0.819 | | |
| 34 | 1 | 1 | 1 | 0.9975 | 1 | 0.9135 | 1 | 0.9135 | 1 | 0.9135 | 1 | 0.9135 | 1 | 0.9135 | 1 | 0.9135 | 1 | 0.9156 | 1 | 1 |
| 35 | 1 | 1 | 1 | 1 | 1 | 0.9961 | 1 | 0.909 | 1 | 0.909 | 1 | 0.909 | 1 | 0.909 | 1 | 0.909 | 2 | 0.909 | 3 | 0.9113 |
| 36 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9957 | 1 | 0.9957 | 1 | 0.9957 | 1 | 0.9957 | 1 | 0.9957 | 1 | 0.9957 | 2 | 0.9959 |

**S8 Table. Minimum number of connected components and the fractional size of the largest component for metabolisms viable on multiple carbon sources.** Each pair of columns shows the minimum number of connected components ($n_C$) as well as the fractional size of the largest component ($r_G$) for metabolism of different sizes n (left-most column) that are viable on k carbon sources (first row). For metabolisms with $n \geq 37$, for every k, $n_C$ and $r_G$ both equal to one. Empty fields mean that no viable metabolism exists for the corresponding k and n.

.

| Abbreviation | Metabolite full name |
|---|---|
| 13dpg | 3-Phospho-D-glyceroyl phosphate |
| 2pg | D-Glycerate 2-phosphate |
| 3pg | 3-Phospho-D-glycerate |
| 6pgc | 6-Phospho-D-gluconate |
| 6pgl | 6-phospho-D-glucono-1,5-lactone |
| ac | Acetate |
| ac[e] | Acetate (extracellular) |
| acald | Acetaldehyde |
| acald[e] | Acetaldehyde (extracellular) |
| accoa | Acetyl-CoA |
| actp | Acetyl phosphate |
| adp | ADP |
| akg | 2-Oxoglutarate |
| akg[e] | 2-Oxoglutarate (extracellular) |
| amp | AMP |
| atp | ATP |
| cit | Citrate |
| co2 | CO2 |
| co2[e] | CO2 (extracellular) |
| coa | Coenzyme A |
| dhap | Dihydroxyacetone phosphate |
| e4p | D-Erythrose 4-phosphate |
| f6p | D-Fructose 6-phosphate |
| fdp | D-Fructose 1,6-bisphosphate |
| for | Formate |
| for[e] | Formate (extracellular) |
| fru[e] | D-Fructose (extracellular) |
| fum | Fumarate |
| fum[e] | Fumarate (extracellular) |
| g3p | Glyceraldehyde 3-phosphate |
| g6p | D-Glucose 6-phosphate |
| glc-D[e] | D-Glucose (extracellular) |
| gln-L | L-Glutamine |
| gln-L[e] | L-Glutamine (extracellular) |
| glu-L | L-Glutamate |
| glu-L[e] | L-Glutamate (extracellular) |
| glu-L[e] | L-Glutamate (extracellular) |
| glx | Glyoxylate |
| h | H+ |
| h[e] | H+ (extracellular) |
| h2o | H2O |
| h2o[e] | H2O (extracellular) |
| icit | Isocitrate |
| lac-D | D-Lactate |
| lac-D[e] | D-Lactate (extracellular) |
| mal-L | L-Malate |
| mal-L[e] | L-Malate (extracellular) |
| nad | Nicotinamide adenine dinucleotide |
| nadh | Nicotinamide adenine dinucleotide (reduced) |
| nadp | Nicotinamide adenine dinucleotide phosphate |
| nadph | Nicotinamide adenine dinucleotide phosphate (reduced) |
| nh4 | Ammonium |
| nh4[e] | Ammonium (extracellular) |
| o2 | O2 |
| o2[e] | O2 (extracellular) |

| | |
|---|---|
| oaa | Oxaloacetate |
| pep | Phosphoenolpyruvate |
| pi | Phosphate |
| pi[e] | Phosphate (extracellular) |
| pyr | Pyruvate |
| pyr[e] | Pyruvate (extracellular) |
| q8 | Ubiquinone-8 |
| q8h2 | Ubiquinol-8 |
| r5p | alpha-D-Ribose 5-phosphate |
| ru5p-D | D-Ribulose 5-phosphate |
| s7p | Sedoheptulose 7-phosphate |
| succ | Succinate |
| succ[e] | Succinate (extracellular) |
| succoa | Succinyl-CoA |
| xu5p-D | D-Xylulose 5-phosphate |

**S9 Table. Metabolites and reactions in central carbon metabolism. Metabolites in central carbon metabolism.** Metabolites abbreviations (left columns) and their full names (right columns) are shown. Note Rows in red correspond to biomass precursors in the central carbon metabolism. atp, nadph and nad are also biomass precursors, but we wish to emphasize on metabolites that are act as biochemical precursors to the actual biomass precursors of *E. coli* (See main text). atp, nadph and nad are also biomass precursors for *E. coli*.

# Chapter 7:

# The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism

Sayed-Rzgar Hosseini and Andreas Wagner

The content of this chapter has been published as:

Hosseini, S.-R., and A. Wagner. 2016. The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. BMC Syst. Biol. 10: 97. https://doi.org/10.1186/s12918-016-0343-7

## 7.1. Abstract

Biological systems are rife with examples of pre-adaptations or exaptations. They range from the molecular scale – lens crystallins, which originated from metabolic enzymes – to the macroscopic scale, such as feathers used in flying, which originally served thermal insulation or waterproofing. An important class of exaptations is novel and useful traits with non-adaptive origins. Whether such origins could be frequent cannot be answered with individual examples, because it is a question about a biological system's *potential* for exaptation.

We here take a step towards answering this question by analyzing central carbon metabolism, and novel traits that allow an organism to survive on novel sources of carbon and energy. We have previously applied flux balance analysis to this system and predicted the viability of $10^{15}$ metabolic genotypes on each of 10 different carbon sources.

We here use this exhaustive genotype-phenotype map to ask whether a central carbon metabolism that is viable on a given, focal carbon source $C$ – the equivalent of an adaptation in our framework – is usually or rarely viable on one or more other carbon sources $C_{new}$ – a potential exaptation. We show that most metabolic genotypes harbor potential exaptations, that is, they are viable on one or more carbon sources $C_{new}$. The nature and number of these carbon sources depends on the focal carbon source $C$ itself, and on the biochemical similarity between $C$ and $C_{new}$. Moreover, metabolisms that show a higher biomass yield on $C$, and that are more complex, i.e., they harbor more metabolic reactions, are viable on a greater number of carbon sources $C_{new}$.

A high potential for exaptation results from correlations between the phenotypes of different genotypes, and such correlations are frequent in central carbon metabolism. If they are similarly abundant in other metabolic or biological systems, innovations may frequently have non-adaptive ("exaptive") origins.

## 7.2. Introduction

One of the most fundamental questions in evolutionary biology regards the origin of qualitatively new and beneficial traits, i.e., evolutionary innovations [1]. On the one hand, such traits can originate as adaptations that help an organism survive or reproduce. On the other hand, they can also have non-adaptive origins as pre-adaptations or exaptations [2, 3]. Darwin was the first to pay attention to the importance of pre-adaptation when he said that "an organ originally constructed for one purpose… may be converted to one for a widely different purpose" [4]. Later on, multiple lines of evidence from the organismal to the molecular scale confirmed the importance of exaptations as important sources of evolutionary innovation [5–7]. A textbook example involves feathers, which are made of keratins, the same proteins that constitute the scales of reptiles. Feathers most likely originally served as thermoregulation and waterproofing, and were only later "exapted" for flying [2]. Many crystallins, light-refracting proteins in eye lenses, originated as metabolic enzymes [8]. More generally, many genes have been coopted into various developmental and physiological functions by changing their patterns of regulation [5]. For example, the Hedgehog signaling protein, responsible for proper limb development in mammals, has also been coopted to paint eyespots in butterflies, and it helps shape feathers in birds [7]. Exaptations may also have been important in human evolution [9].

It is easy to find examples of exaptations, but much more difficult to find out how frequently any one biological system can bring forth non-adaptive traits that could turn into exaptations. This is not a question about natural history, but about a biological system's *potential* for exaptation. It is the focus of this contribution. The question can only be answered in systems where one can study, either experimentally or computationally, many genotypes and the phenotypes that they form [10–14]. In doing so, one can ask whether a beneficial phenotypic trait frequently entails other traits with the potential to become an exaptation.

Metabolism is a well-suited system for this purpose, and for two main reasons. First, metabolism is a source of multiple evolutionary innovations, especially in microorganisms. For example, microorganisms have acquired the ability to extract energy from non-natural substances, including toxic compounds [15–18]. By

producing novel molecules such as ectoine or glycine betaine, halophilic bacteria can tolerate high salt concentrations [19]. And microbial isolates from pristine soils show not only resistance to a wide range of antibiotics, but many of them are also capable of using these molecules as sources of energy and chemical elements [20].

Second, one can predict novel metabolic phenotypes using computational tools such as flux balance analysis (FBA) for multiple metabolic genotypes [21–26]. The metabolic genotype of an organism is a string of DNA encoding the enzymes catalyzing metabolic reactions, but for computational expediency, a more compact representation of a metabolic genotype based on reactions rather than genes is often used [21–24, 27, 28]. Specifically, given a known "universe" of enzyme-catalyzed biochemical reactions, one can represent the metabolic genotype of an organism as a binary vector whose *i*-th entry corresponds to the *i*-th reaction in this reaction universe [29]. If an organism's genome encodes an enzyme capable of catalyzing a given reaction, the corresponding entry in the genotype vector will be one and zero otherwise. The collection of all such vectors constitutes a metabolic genotype space, and any one organism's metabolic genotype can be thought of as a point in this space. FBA can predict metabolic phenotypes, such as viability (the ability of a metabolism to sustain life in a given spectrum of chemical environments) for any one metabolic genotype. Importantly, FBA-based predictions of viability are in good agreement with experimental data [27, 30–35].

In previous work, we analyzed potential exaptations in genome-scale metabolisms [36]. This work relied on sampling of metabolic genotypes from a vast metabolic genotype space [14, 26]. Because any such sample represents a tiny fraction of the whole space, we here complement this analysis with a more comprehensive approach that examines all members of a genotype space. This is impossible for genome-scale metabolisms, because of their astronomical numbers, but it is possible for smaller-scale metabolic systems, such as a genotype space defined by the 51 biochemical reactions of central carbon metabolism [29].

Central carbon metabolism is a small but crucial part of metabolism, because it plays a pivotal role in life by extracting energy from extracellular carbon sources (Additional files 1 and 2). It includes the interrelated biochemical pathways of glycolysis, gluconeogenesis, the pentose-phosphate pathway (PP), and the tricarboxylic acid cycle (TCA), which are supplemented by anaplerotic reactions and

the glyoxylate shunt [37]. Glycolysis creates high-energy compounds like ATP and NADH and converts glucose into pyruvate. The tricarboxylic acid cycle (TCA) generates ATP, NADH, and amino acid precursors from acetyl-CoA, which results from oxidation of the glycolytic end product pyruvate. The pentose-phosphate pathway produces NADPH and pentose sugars for anabolic reactions. Finally, the reactions of the oxidative phosphorylation pathway participate in production of ATP from NADH.

We here use the genotype-phenotype map of central carbon metabolism to ask how often metabolisms viable on a given carbon source $C$ can survive on one or more other carbon sources $C_{new}$. We show that this is the case for most metabolisms, and we analyze which properties of a metabolism facilitate its potential for exaptation. These properties include the complexity of a metabolism and its efficiency in converting nutrients into biomass. We emphasize that we are not focused on the evolutionary history of central carbon metabolism, but on the potential for its biochemical pathways to bring forth exaptations.

## 7.3. Results

### The genotype space of central carbon metabolism

The genotype space we consider includes all $2^{51} \approx 10^{15}$ metabolic genotypes whose reactions form a subset of the 51 internal reactions of the central carbon metabolism of *E. coli* [29]. Each genotype specifies a chemical reaction network that we refer to as a central carbon metabolism. We call a genotype (metabolism) viable on a given carbon source, if it can synthesize each one of 13 biomass precursors from this source in an otherwise minimal chemical environment (Additional files 1 and 2) [38]. In previous work, we determined the viability of each of the $2^{51}$ genotypes on 10 different carbon sources [39–41], and found that only a tiny fraction of genotypes can sustain life on any one carbon source. This fraction ranges from $10^{-8}$ (on acetate) to $10^{-6}$ (on glucose), corresponding to between $\approx 10^{7}$ and $\approx 10^{9}$ genotypes that are viable on acetate and glucose, respectively. Genotypes viable on a given carbon source form a connected network in genotype space, which implies that different metabolisms can be converted into each other in few viability-preserving mutational steps [39]. We use *E. coli* central metabolism as a departure point for our analysis for two reason. First, it is small enough to be amenable to exhaustive genotype-phenotype mapping, yet

large enough to show multiple different phenotypes. Second, *E. coli* central carbon metabolism is especially well characterized and reasonably complete [38, 42]. Other genotypes in the genotype space we examine correspond to incomplete variants, such as those lacking a complete citric acid cycle or having incomplete pentose-phosphate pathway.

The fundamental question we ask here is whether a metabolism that is viable on a specific focal carbon source $C$ is usually also viable on one or more other carbon sources $C_{new}$, which corresponds to a potential exaptation or preadaptation in the framework of our computational analysis. In a previous analysis, we had asked this question for randomly sampled genome-scale metabolisms required to be viable on a specific carbon source $C$ [36], and we here extend this approach to all $\approx 10^{15}$ central carbon metabolisms whose phenotypes we have previously exhaustively enumerated [39–41].

**High potential for exaptation in central carbon metabolism**

We first defined an exaptation index $I$ as the number of carbon sources $C_{new}$ on which a metabolism is viable (in addition to the carbon source $C$) [36]. We then asked what fraction of metabolisms viable on $C$ have $I>0$, i.e., they are exapted to at least one additional carbon source. Figure 1A shows that for all ten focal carbon sources $C$ we consider here except one, the majority of metabolisms are viable on at least one carbon source $C_{new}$. For example, 95 percent of metabolisms viable on glucose are also viable on at least one additional carbon source. The one exception is α-ketoglutarate for which only 38 percent of viable metabolisms are also viable on additional carbon sources (Figure 1A). Another extreme is represented by metabolisms viable on fructose and fumarate, all of which are viable on additional carbon sources. The reason is that all metabolisms viable on fructose are also viable on glucose, and all metabolisms viable on fumarate are also viable on succinate. In sum, central carbon metabolism harbors great potential for exaptation.

One can partition metabolic genotype space according to the complexity or size of genotypes, defined as the number of reactions $n$ in a metabolism [40]. Any one metabolism needs to have a minimal size $n$ for viability, which depends on the carbon source considered, and ranges from $n=23$ for glucose to $n=30$ for acetate. We next

asked if the fraction of metabolisms with $I > 0$ depends on this metabolism size. Figure 1B shows that it does. For any one carbon source $C$, more complex metabolisms have a higher potential for exaptation. The exceptions are metabolisms viable on fumarate and fructose, because all of them have $I>0$, regardless of size.

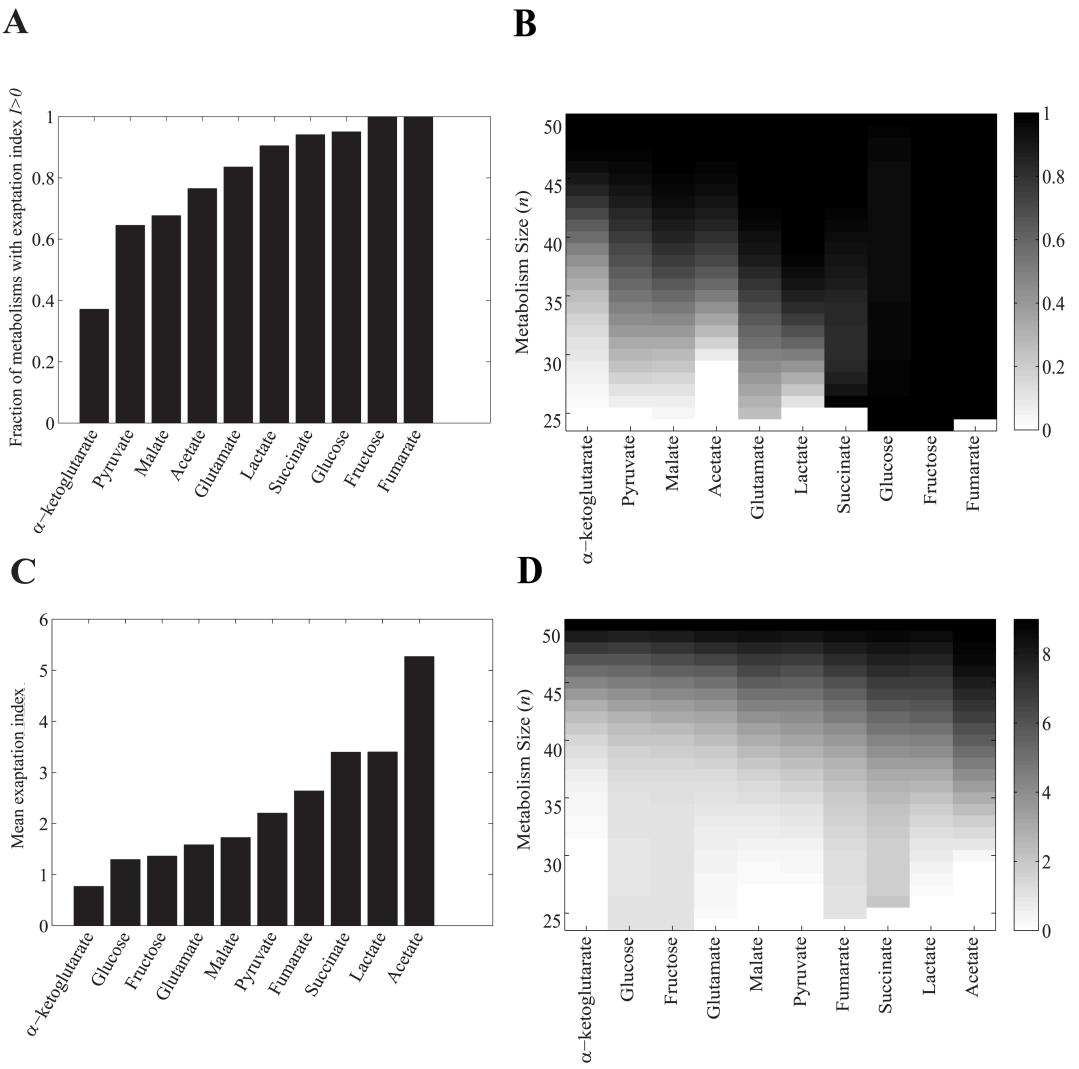



**Figure 1 High exaptation potential in central carbon metabolisms. (A)** Fraction of metabolisms with exaptation index $I>0$ ($y$-axis) viable on some carbon source $C$ ($x$-axis); **(B)** fraction of metabolisms (coded by shade of grey, see legend) with exaptation index ($I>0$) that are viable on some carbon source $C$ ($x$-axis) and have a given size ($y$-axis), **(C)** mean exaptation index of metabolisms viable on some carbon source $C$ ($x$-axis), and **(D)** mean exaptation index (coded by shade of grey, see legend) of metabolisms viable on some carbon source $C$ ($x$-axis) and with a given size ($y$-axis). White in (B) and (D) corresponds to metabolisms whose size is too small for viability on $C$.

We next determined the average exaptation index among all viable metabolisms, which indicates the average number of additional carbon sources $C_{new}$ that a metabolism viable on some carbon source $C$ is also viable on. This average exaptation index exceeds 1 in all of the carbon sources except for α-ketoglutarate, where it is 0.88 (Figure 1C). For acetate, this index has the largest value ($I=5.27$), which implies that a metabolism viable on acetate will, on average, be viable on more than 5 of the remaining 9 carbon sources. The index also increases with increasing metabolism size (Figure 1D), meaning that larger metabolisms are viable on more carbon sources $C_{new}$. Furthermore, we observed that the exaptation index varies widely among metabolisms of the same size and viable on the same carbon source (Additional file 3).

In a final analysis, we asked whether high exaptation potential might be caused preferentially by reactions that are disconnected from the rest of metabolism. Such disconnected reactions must fulfill at least one of the following two criteria. First, their products are neither biomass precursors nor substrates of any other reaction of a given metabolism. Second, at least one of their substrates is neither a product of other reactions nor a nutrient taken up from the environment. To find out how disconnected reactions affect exaptation potential, we eliminated from our analysis those metabolisms harboring such reactions, and observed that the exaptation indices remain almost unchanged (Additional file 4). The incidence of disconnected reactions does not strongly affect the exaptation potential of central carbon metabolism.

In sum, viability on a given carbon source generally entails viability on other carbon sources, and often on multiple such carbon sources. Thus, central carbon metabolism has a high potential for exaptation. This potential increases with metabolic complexity, i.e., with the number of reactions in a metabolism.

**Minimal central carbon metabolisms also harbor exaptation potential**

A special role in our analysis is played by metabolisms that are *minimal*. We define them as metabolisms from which not a single reaction can be removed without abolishing viability on the focal carbon source. We note that there may be multiple such metabolisms, and that they are not necessarily the smallest possible metabolisms viable on this carbon sources. They are important, because they harbor only essential

reactions. If exaptation potential depends on non-essential reactions, then it is possible that such minimal metabolisms harbor no exaptation potential.

To find out, we first identified all minimal central carbon metabolisms viable on a given focal carbon source (table 1). For example, 161 minimal central carbon metabolisms are viable on glucose, and their size varies from 23 to 30 reactions [39].

| **Focal Carbon Source** | **Number of Minimal Metabolisms** | **Size Range** | **Number (percentage) of Minimal Metabolisms with *I>0*** |
|---|---|---|---|
| Fructose | 146 | [23-30] | 146 (100.00%) |
| Fumarate | 456 | [26-32] | 456 (100.00%) |
| Glucose | 161 | [23-30] | 154 (95.65%) |
| Succinate | 348 | [27-32] | 304 (87.36%) |
| Lactate | 180 | [26-34] | 144 (80.00%) |
| Malate | 456 | [25-32] | 176 (38.60%) |
| Glutamate | 187 | [26-32] | 72 (38.50%) |
| Acetate | 76 | [30-36] | 16 (21.05%) |
| Pyruvate | 569 | [26-34] | 96 (16.87%) |
| α-ketoglutarate | 970 | [25-32] | 80 (8.25%) |

**Table 1: Exaptation potential of minimal metabolisms.** Columns, from left to right, indicate the focal carbon source, the number of minimal metabolisms that are viable on this carbon source, the size range of these metabolisms, and the number (percentage) of minimal metabolisms with exaptation index (*I>0*).

The ten focal carbon sources in Table 1 can be subdivided into two groups. In the first group (fructose, fumarate, glucose, succinate and lactate), the vast majority (80-100 percent) of minimal central carbon metabolisms have exaptation indices *I>0*. For example, among the 161 minimal metabolisms viable on glucose, 154 (95.7%) can survive on at least one additional carbon source (146 on one, and eight on two additional carbon sources). Three of these 154 minimal metabolisms have the smallest possible size for metabolisms viable on glucose (*n=23* reactions), and each of these three is viable on one additional carbon source. For the second group of focal carbon sources (table 1, malate, glutamate, acetate, pyruvate, and α-ketoglutarate), fewer than 50 percent of minimal metabolisms show an exaptation index *I*>0. For example, among the 76 minimal metabolisms viable on acetate, 60 are only viable on acetate and only 16 (21%) of them can survive on another carbon source. In sum, there are clear differences among carbon sources in the exaptation potential of minimal metabolisms. However, on all carbon sources some minimal metabolisms

show exaptation potential, and on half of the carbon sources the vast majority of minimal metabolisms does. Non-essential reactions are not solely responsible for the exaptation potential of central carbon metabolisms.

That being said, reactions that are non-essential on any one carbon source do play a role in increasing a metabolism's exaptation potential, but the importance of this role depends on the carbon source. We demonstrated this with the following approach, applied to all carbon sources, all minimal metabolisms for each carbon source, and all possible numbers $n_{ne}$ of non-essential reactions. We identified all $n_{ne}$ -tuples of such reactions, i.e., reactions that are not already part of the metabolism, added each $n_{ne}$-tuple to the minimal metabolism, and determined the exaptation index $I$ of the resulting metabolism. Figure 2A shows the number of added non-essential reactions together with the fraction of metabolisms with an exaptation index $I>0$, for two representative carbon sources from the two groups, glucose (group 1) and acetate (group 2). For glucose, where most minimal metabolisms already have exaptation potential, adding non-essential reactions cannot strongly increase this potential. Specifically, the fraction of metabolisms with exaptation index ($I>0$) grows very slowly and it does not reach one even after addition of 20 non-essential reactions to some of the minimal metabolisms. Figure 2B shows the exaptation index itself as a function of the number $n_{ne}$ of added reactions. For glucose, it remains nearly unchanged after adding 5 non-essential reactions to minimal networks, and starts to increase only thereafter. In contrast, for acetate, where the fraction of metabolisms with $I>0$ is low for $n_{ne}=0$, this fraction rises rapidly, to over 60 percent after adding 5 reactions, and to 100 percent after adding 17 reactions (Figure 2A). Moreover, the exaptation index itself increases rapidly. It surpasses the exaptation index of minimal metabolisms viable on glucose after adding merely three non-essential reactions, and increases rapidly thereafter as well. In sum, even minimal metabolisms have some exaptation potential, and in those with low exaptation potential, the addition of non-essential reactions increases this potential to the greatest extent.

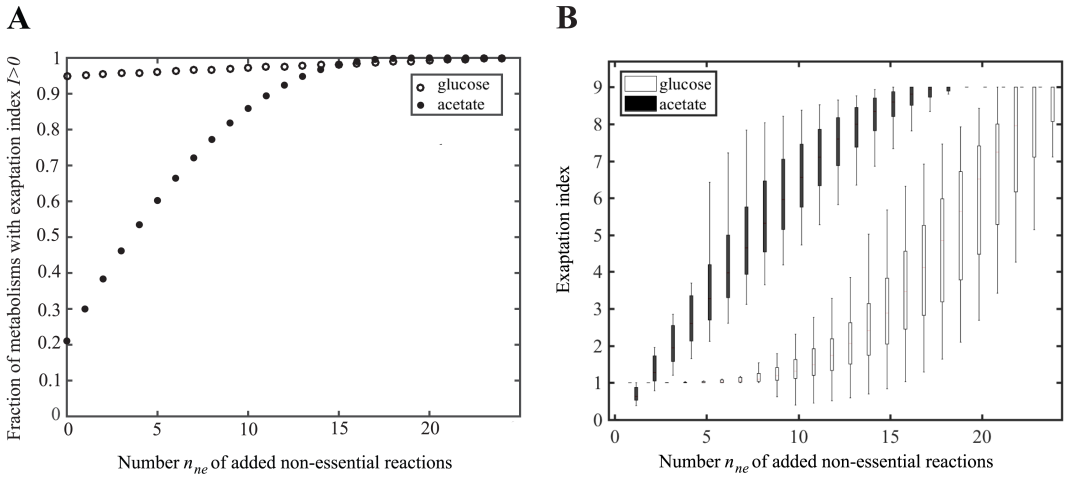



**Figure 2 Exaptation potential and non-essential reactions.** Vertical axes show **(A)** the fraction of metabolisms with exaptation index (*I>0*), and **(B)** the exaptation index itself, among metabolisms generated by adding a given number $n_{ne}$ of non-essential reactions (*x*-axis) to the minimal metabolisms viable on glucose (open circles/boxes), and acetate (filled circles/boxes). Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. Note that this analysis is exhaustive, meaning that (*i*) all minimal metabolisms viable on glucose (161), and acetate (76) are considered, and (*ii*) all possible $n_{ne}$ -tuples of non-essential reactions (*x*-axis) have been added to each minimal metabolism.

## Metabolisms viable on a given focal carbon source can be pre-adapted to a wide variety of other carbon sources

Our analysis thus far has not addressed the question whether *different* metabolisms viable on some carbon source *C* show exaptation to *different* carbon sources $C_{new}$. To answer this question, we separated metabolisms according to their focal carbon source *C* and their numbers of reactions, and determined the number of carbon sources $C_{new}$ on which metabolisms in each of these categories are viable. For all except the smallest metabolisms (*n<30*), at least one metabolism in each category is viable on each of the nine possible carbon sources $C_{new}$ (Additional file 5). In other words, regardless of the focal carbon source *C*, exaptation is possible on every single alternative carbon source.

For the next step of our analysis, we represented viability on each of the nine carbon sources $C_{new}$ as a binary phenotype vector. For any one metabolism, this vector contains a one for each carbon source $C_{new}$, on which that metabolism is viable, and a zero otherwise. We defined the phenotypic distance between a given pair of metabolisms as the Hamming distance between these phenotype vectors. The greater

this distance is for two metabolisms, the larger is the number of carbon sources $C_{new}$ on which one metabolism is viable and the other is not.

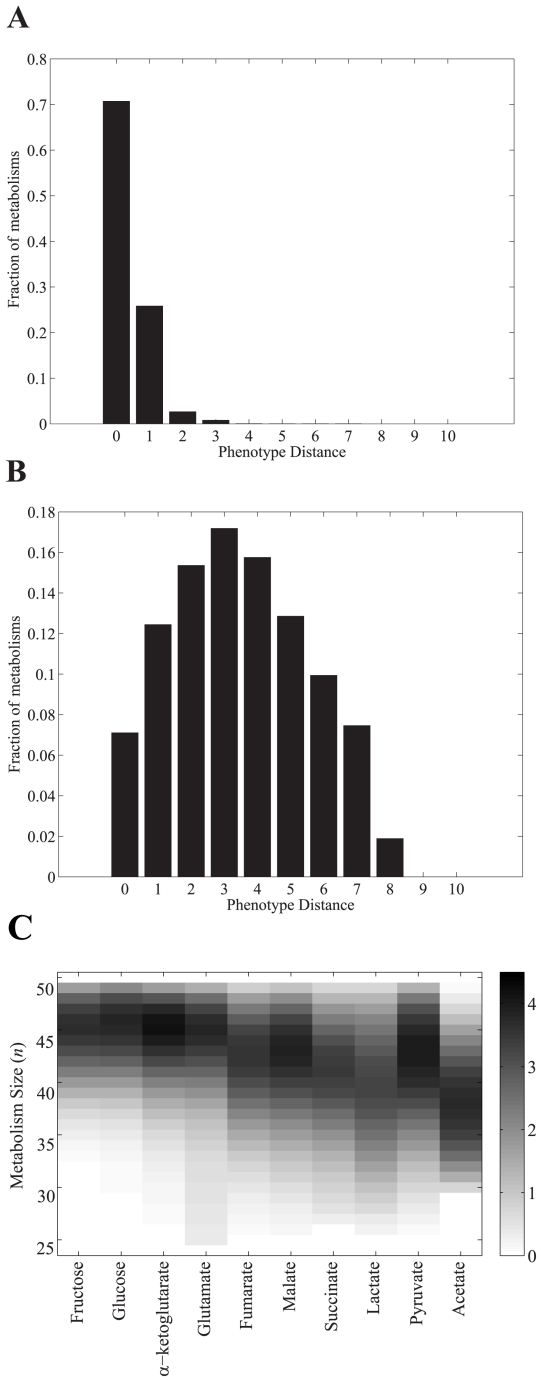



**Figure 3 Metabolisms can preadapt to a wide variety of carbon sources.** Panels **(A)** and **(B)** show histogram of the phenotype distance (*x*-axis), for metabolisms of size (A) 30, and (B) 45 viable on glucose as carbon source *C*. **(C)** Mean phenotypic distance (coded by shade of grey, see legend) of metabolisms viable on a focal carbon source (*x*-axis) and with a given number of reactions *n* (*y*-axis

Figures 3A and 3B show examples of the distribution of the phenotype distance for metabolisms of *n=30* and *n=45* reactions viable on the focal carbon source glucose. 29.28% (*n=30*) and 92.89% (*n=45*) of all metabolism pairs have phenotypic distance larger than or equal to one, and 32.15% (*n=45*) of all metabolism pairs have phenotypic distance larger than or equal to five. Phenotypic distances can reach

values up to eight, meaning that two metabolisms may share viability on only one of the nine possible carbon sources $C_{new}$. (See also Additional file 6 for the remaining glucose panels and Additional file 7, where $C$ is pyruvate). Figure 3C shows the mean phenotypic distances for metabolisms of different sizes $n$ and focal carbon sources $C$. It illustrates that large phenotypic distances are not peculiarities of metabolism pairs viable on glucose or pyruvate. Also, for each focal carbon source $C$, phenotypic distance generally increases with metabolism size. The only exception involves the largest metabolisms ($n>48$), where the average phenotypic distance is low and decreases with increasing $n$. The reason is that the largest metabolisms are highly likely to be viable on all ten carbon sources, which lowers their phenotypic distance. Similar observations hold for metabolisms without disconnected reactions (Additional files 8, 9 and 10).

In sum, different metabolisms viable on a given carbon source are usually exapted to different additional carbon sources, and this exaptive diversity increases with a metabolism's size. The exaptation potential of central carbon metabolism can give rise to multiple different metabolic innovations.

**The potential for pre-adaptation depends on the biochemical similarity between carbon sources**

In a next analysis, we asked whether different carbon sources $C_{new}$ are equally likely to occur as exaptations. Figure 4A shows, for each of nine carbon sources $C_{new}$, and for metabolisms whose focal carbon source $C$ is glucose, the fraction of metabolisms that are also viable on $C_{new}$. The figure indicates huge disparities between carbon sources, where 97.5% of metabolisms viable on glucose are also viable on fructose, but only 11.7% are viable on malate, and fewer than 10% are viable on any of the other 7 carbon sources. Figure 4B shows these fractions broken down by metabolism size $n$. Almost all metabolisms viable on glucose are also viable on fructose, regardless of $n$, but the potential for preadaptation to other carbon sources is monotonically increasing with increasing $n$. Note that the only metabolism with the maximum of $n$=51 reactions is viable on all 10 carbon sources, such that at the highest $n$, the potential for preadaptation to any carbon source must reach one. These patterns are not a peculiarity of metabolisms viable on glucose, as Additional file 11 illustrates for metabolisms with lactate as the focal carbon source $C$.

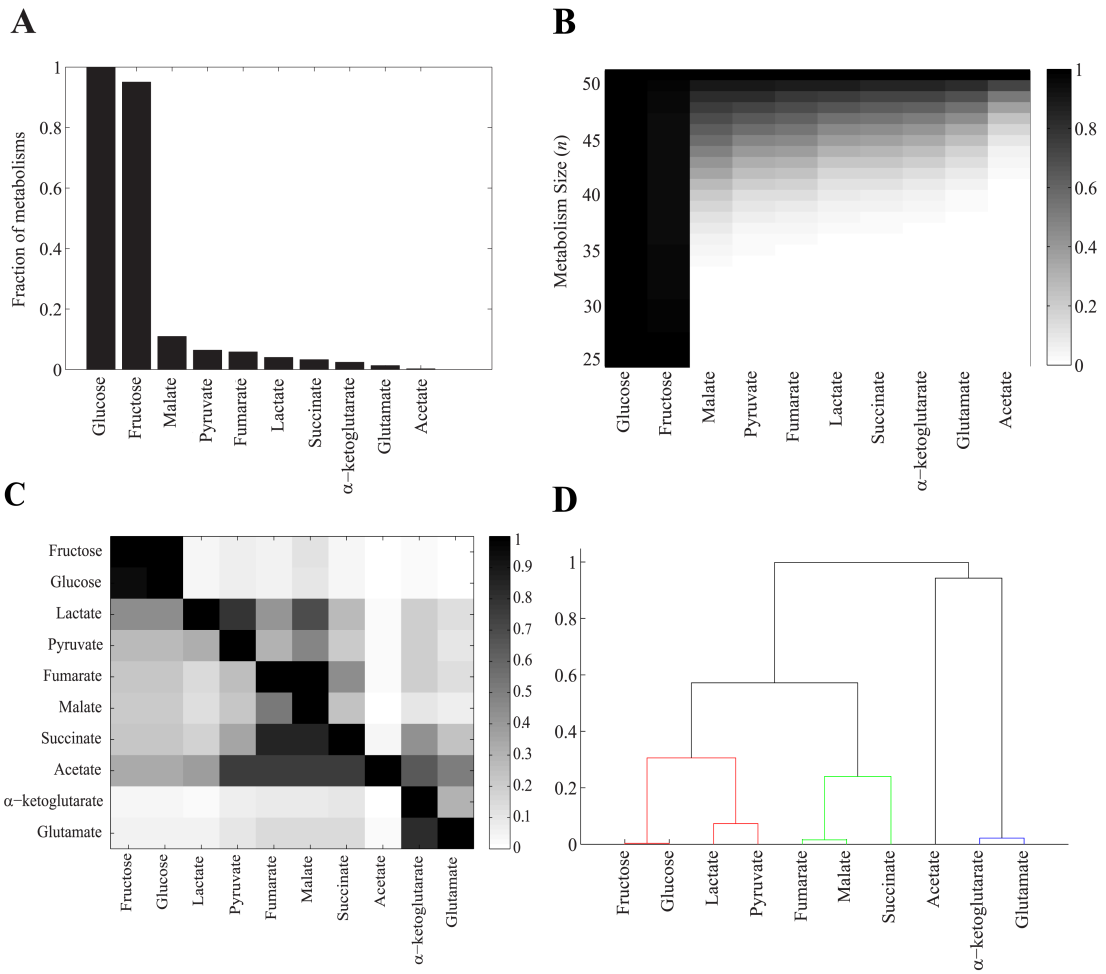



**Figure 4 Potential for preadaptation depends on biochemical similarity between carbon sources.** **(A)** The histogram shows the fraction of metabolisms viable on glucose as carbon source *C* that are also viable on each of the nine other carbon sources $C_{new}$ (*x*-axis). **(B)** As in (A), but broken down by metabolism size, and fractions of viable metabolisms are represented by different shades of grey (see legend). **(C)** Fraction of metabolisms viable on carbon source *C* (*x*-axis), which are also viable on carbon source $C_{new}$ (*y*-axis), (coded by shade of grey, see legend). **(D)** Dendrogram of carbon sources clustered based on their pairwise preadaptation propensity. We used UPGMA method (unweighted pair group method with arithmetic means), for clustering carbon sources.

Figure 4C extends this analysis to all carbon sources *C*. Its *x*-axis shows the focal carbon source *C*, its *y*-axis the carbon source $C_{new}$, and the grey shading of each matrix entry corresponds to the fraction of metabolisms viable on *C* and $C_{new}$. Importantly, this matrix is not symmetric, showing that the potential of preadaptation between a given pair of carbon sources is not necessarily reciprocal. For example, the probability of preadaptation to glucose among metabolisms viable on acetate is 0.33 but the probability of preadaptation to acetate among metabolisms viable on glucose is only 0.0023. Moreover, all the metabolisms viable on fumarate are also viable on

malate (i.e. *preadaptation probability=1*), but only 52.73% of metabolisms viable on malate are also viable on fumarate (i.e. *preadaptation probability =0.5273*). This asymmetry comes from the relative position of carbon sources in central carbon metabolism. For example, fumarate precedes malate in the citric acid cycle (Additional files 1 and 2), because malate is synthesized from fumarate. This ordering means that metabolisms viable on fumarate will also frequently be viable on malate, whereas the opposite is not necessarily true.

In a final analysis, we also clustered carbon sources according to their mutual propensity for preadaptation, using the hierarchical clustering method UPGMA (unweighted pair group method with arithmetic means) [43]. Figure 4D shows the resulting dendrogram. The carbon sources that cluster together are biochemically closely related, which can help explain their mutual propensity for pre-adaptation. Specifically, (i) glucose and fructose are both glycolytic carbon sources, (ii) fumarate, succinate and malate occupy consecutive steps in the citric acid cycle, (iii) pyruvate and lactate are interconvertible via lactate dehydrogenase, (iii) acetate is functionally linked to pyruvate via acetyl-coenzymeA, which is produced from pyruvate through a reaction catalyzed by pyruvate dehydrogenase, and (iv) glutamate and α-ketoglutarate are interconvertible via glutamate dehydrogenase. An analysis of metabolisms without disconnected reactions shows a similar pattern (Additional files 12 and 13). In sum, metabolisms viable on biochemically similar focal carbon sources $C$ also tend to be pre-adapted to biochemically similar carbon sources $C_{new}$.

**High biomass yield and low waste production are associated with greater potential for pre-adaptation**

Metabolisms of the same size and that are viable on the same carbon source $C$ can vary widely in their exaptation index, i.e., the number of additional carbon sources $C_{new}$ on which they are viable. To understand the causes of this variation, we analyzed the average biomass yield per mole of carbon. We found that this yield increases with increasing exaptation index, regardless of the focal carbon source $C$ (Figure 5A), and regardless of metabolism size (Additional file 14). We also examined the number of waste metabolites, molecules that a metabolism synthesizes (and excretes) but that are not biomass precursors. This number of molecules can

vary widely for metabolisms of the same size that are viable on the same carbon source *C* (Figure 5B, and Additional file 15). Not surprisingly, metabolisms that show higher yield also excrete fewer waste molecules, regardless of their size, and regardless of their focal carbon source *C* (Figure 5C and Additional file 16). In addition, we observed that metabolisms with more reactions generally produce less waste, regardless of their focal carbon source (Figure 5D).

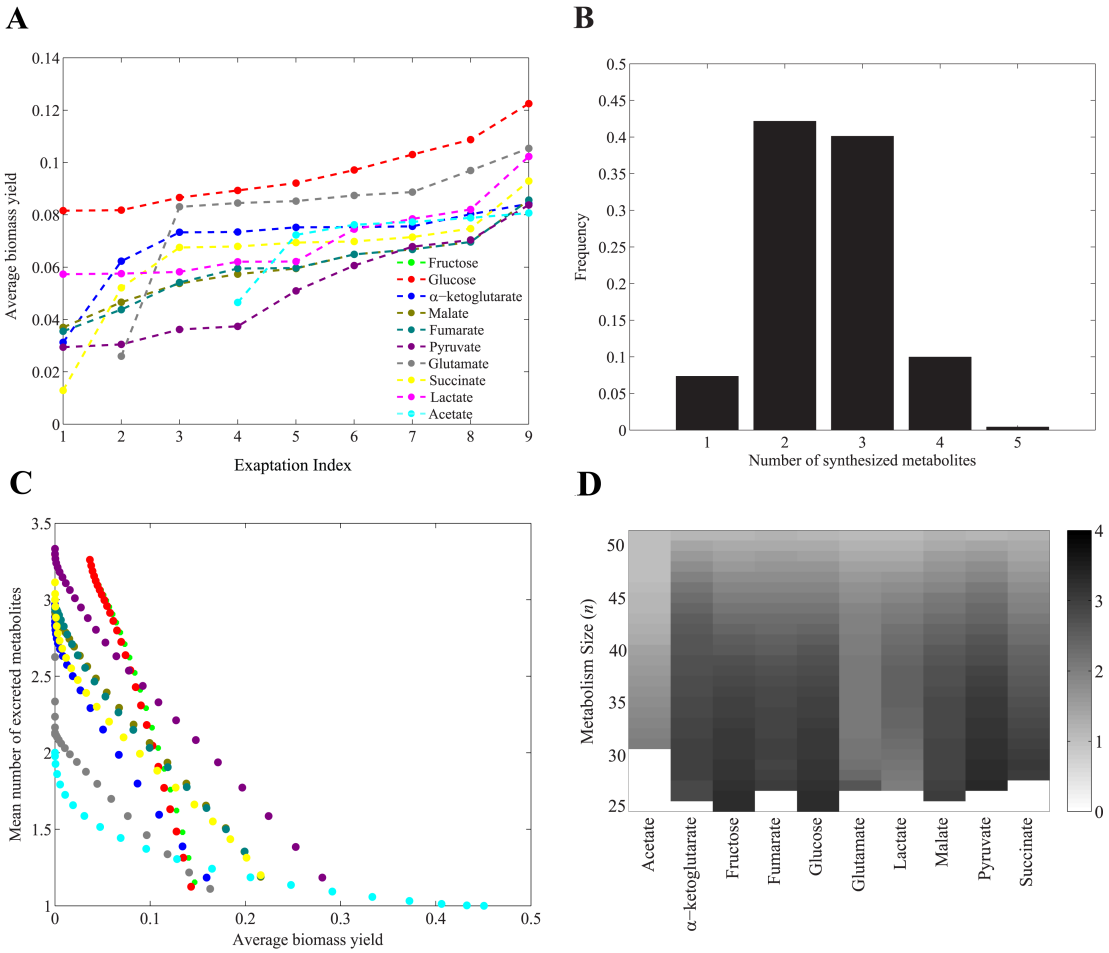



**Figure 5 High biomass yield and low waste production are associated with greater potential for preadaptation. (A)** The *x*-axis shows the exaptation index, i.e., the number of carbon sources $C_{new}$ on which metabolisms viable on carbon source *C* (color legend) are viable. The *y*-axes show the average biomass yield. Data is based on metabolisms of size *n=40*. **(B)** Fraction of metabolisms excreting a given number of metabolites (*x*-axis) among metabolisms of size 40 and viable on glucose. **(C)** Each point shows mean number of excreted metabolites (*y*-axis), and mean biomass (*x*-axis) among metabolisms of a given size viable on a given carbon source colored according to legend in (A). **(D)** Mean of the number of excreted metabolites (coded by shade of grey, see legend) among metabolisms of a given size (*y*-axis), that are viable on a given carbon source (*x*-axis). White colors correspond to metabolisms whose size is too small for viability on *C*.

Taken together, these observations show that metabolisms with higher exaptation index are more efficient, converting more of their carbon source *C* into biomass, and excreting fewer waste products. Larger metabolisms synthesize less waste and show higher biomass yield (Additional file 17), which can help explain their greater potential for exaptation (Figure 1B and 1D).

## 7.4. Discussion

We systematically analyzed the potential for exaptation or preadaptation in the simple but biologically important system of central carbon metabolism. Our analysis complements a previous study based on sampling a much larger metabolic genotype space [36], because it examines the potential for exaptation in an exhaustively enumerable genotype space of $10^{15}$ metabolisms. Our observations are consistent with that of the previous study, in that we also find a high potential for exaptation in this smaller metabolic system.

Our central observation is that most metabolisms viable on a given carbon source *C* are also viable on one additional carbon source $C_{new}$, and often on multiple such carbon sources. In a changing chemical environment, where the focal carbon source *C* has been consumed but where one or more carbon sources $C_{new}$ become available, this ability can become an important innovation. In other words, potential exaptations are highly abundant, even in a simple metabolic system. They occur preferentially for carbon sources $C_{new}$ that are biochemically similar to *C*, and are ultimately caused by shared biochemical pathways that connect different extracellular carbon sources to essential biomass precursors. Such shared pathways result in complex phenotypic correlations among different genotypes.

We also observed that different metabolisms viable on a given carbon source *C* can be preadapted to widely different sets of carbon sources $C_{new}$. This diversity of preadaptation can help make populations robust to environmental changes in carbon source availability, because a sufficiently large and genotypically diverse population may harbor at least one metabolic genotype that is already preadapted to some newly available $C_{new}$.

One advantage of an exhaustive enumeration approach like ours is that it allows us to systematically identify metabolic properties associated with high exaptation potential. One such property is metabolic complexity, i.e., the number of reactions in a metabolism. The greater metabolic complexity is, the greater is the number of carbon sources $C_{new}$ on which a metabolism is viable. Another property is metabolic efficiency, the ability to convert carbon into biomass (precursors) with few waste products. The more efficient a metabolism is, the greater is its potential for exaptation. Complexity and efficiency are linked, because at least in our study system, larger metabolisms produce less waste. These associations might be more difficult to understand in genome-scale systems with thousands of reactions, but studying them in a simpler system leads to a straightforward explanation. Specifically, in a larger metabolism it is more likely that most reactions link carbon sources and biomass productively, without producing dead-end products that cannot be used by other reactions. And this very feature makes it also more likely that a reaction path exists from any one carbon source $C_{new}$ to each biomass precursor.

Among the limitations of our study is that we focused on carbon metabolism, and the metabolism of other chemical elements, such as nitrogen or sulfur, may differ in its exaptation potential. To find out whether this is the case remains a task for future work, but we note that sulfur and nitrogen metabolism generally show similar properties to carbon metabolism in studies of metabolic genotype spaces [44, 45].

A second limitation is that our analysis focuses on the presence or absence of reactions, and it neglects regulatory constraints arising through sub-optimal expression or regulation of an enzyme. This is consistent with our focus on the qualitative feature of viability, for which the presence or absence of reactions (enzymes) is more important than their quantitative regulation. We also note that regulatory constraints can be readily broken in evolution, even on the short time scales of laboratory evolution experiments [46–48].

A third limitation comes from our reaction-centered definition of metabolic genotypes. This coarse-grained definition of metabolic genotype is widely used [21–24, 27, 28] , because it is simple, computationally efficient, and yet sufficiently informative for many analyses. However, it neglects that there need not be a one-to-one relationship between metabolic genes and metabolic reactions. Some reactions are catalyzed by multiple enzymes [49], and some enzymes catalyze multiple

reactions [50–52]. One recent study that speaks to this limitation has focused on promiscuous enzymes that catalyze multiple biochemical reactions in genome-scale metabolisms. It shows that considering this one-to-many relationship between enzymes and reactions leads to an increase in the number of environments in which a metabolism is viable [53]. This would also apply to our study system, because adding reactions catalyzed by promiscuous enzymes to a metabolism can only increase its potential for exaptation. That is, addition of a reaction cannot abolish viability on the focal carbon source, but it might convey the ability to survive on further carbon sources $C_{new}$

Fourth, while *E. coli* central carbon metabolism is more complete than that of other species, where, for example, parts of the citric acid cycle are missing [42], we have not considered all reactions that could be considered part of a central carbon metabolism. For example, we have only considered the canonical Embden-Meyerhof-Parnas glycolytic pathway, and we have neglected reactions belonging to the Entner-Doudoroff (ED) and the phosphoketolase pathways [54, 55]. This was necessary, because the size of the genotype space we analyze is already large ($10^{15}$ genotypes), and at the limit of feasibility for exhaustive genotype-phenotype mapping [40, 41]. Any additional reactions would have made an exhaustive analysis impossible. Just as for the preceding limitation, we note that adding these or other reactions to any one metabolism could only increase its potential for exaptation. For this reason, considering more complex metabolisms would not affect our core observation that many variants of central carbon metabolism harbor exaptive potential.

Finally, we only considered 10 carbon sources, whereas metabolic generalists like *E. coli* can be viable on many more carbon sources [56, 57]. However, even this low number of carbon sources was sufficient to detect a high potential for exaptation, and once again, considering more carbon sources could only increase the estimated exaptive potential.

One can envision the following evolutionary scenario in which traits with exaptive potential facilitate survival in novel environments. Consider a minimal metabolic network adapted to a specific carbon source $C$, i.e., a network from which no reaction can be removed without abolishing viability on $C$. Many such minimal networks are also viable on one or more additional carbon sources $C_{new}$. If $C_{new}$ becomes available and the organism hosting this metabolism (or its competitors) has consumed $C$, then

viability on $C_{new}$ becomes an adaptation that helps the organism survive. The survivor's descendants undergo processes such as gene duplication, point mutations, and horizontal gene transfer, which may enable some of them to catalyze novel reactions that allow it to utilize a carbon source *C'* (and as a by-product, perhaps one or more additional carbon sources $C'_{new}$). If none of these carbon sources ever occur in the environment, any such genetic change will eventually disappear through genetic drift or degenerative mutations. However, if *C'* occurs in the environment, a new adaptation with exaptive potential on carbon source $C'_{new}$ has arisen. In other words, one can envision a step-wise expansion of metabolism that is driven by adaptive processes, but in which the exaptive potential of some traits facilitates survival in novel environments. That different genotypes viable on *C* are viable on different carbon sources $C_{new}$ (figure 3) may further facilitate adaptive evolution.

The high exaptation potential of central carbon metabolism, and of genome-scale metabolisms in general [36] invites speculation that many metabolic innovations originate non-adaptively. However, we emphasize that our analysis is not suited to identify any one metabolic trait as an exaptation. It can thus also not identify the incidence of exaptations in metabolic evolution, which remains a major challenge for future work.

We analyzed central carbon metabolism, a metabolic system small enough to lend itself to exhaustive genotype-phenotype mapping, and have systematically quantified this system's potential for preadaptation for viability on novel carbon sources. Our results indicate that metabolisms viable on any one carbon source can be preadapted to multiple other carbon sources. The potential for such preadaptation rises with the complexity of a metabolism, i.e., with its numbers of reactions, and with its efficiency. It results from correlations between the phenotypes of different genotypes, which are caused by shared pathways that connect different extracellular carbon sources to essential biomass precursors

## 7.5. Methods

### Flux balance analysis

Flux Balance Analysis (FBA) is a widely used computational method that predicts the metabolic flux through biochemical reactions in metabolic networks [58–61]. FBA uses information about the stoichiometric coefficients of the metabolites

participating in each reaction, encapsulated in the stoichiometric matrix *S*, which is of dimension $m \times n$, where *m* and *n*, respectively, denote the number of metabolites and the number of reactions in a metabolism. FBA assumes that a metabolism has reached a steady-state, as might be attained by a growing population of bacteria in chemostat with constant nutrient supply, where mass conservation constraints apply. These constraints are mathematically expressed as $Sv = 0$, where *v* is the vector of fluxes ($v_i$) through reaction *i*. The solution space of this equation is called the null space of the stoichiometric matrix (*S*). This null space is further constrained by upper and lower bounds on the fluxes through each reaction. FBA applies linear programming to find the optimal flux vector(s) that maximize an objective function Z. This task can be mathematically formulated as finding a flux vector $(v^*)$ with the property

$v^* = max_v\, Z(v) = \{c^T v | S.v = 0,\ a \leq v \leq b\},$

where the vector *c* contains scalar coefficients representing a maximization criterion, and entries $a_i$ and $b_i$ of vectors $a$ and $b$, respectively, indicate the minimally and maximally possible flux through reaction *i*.

We use a set of 13 well-known precursors from central carbon metabolism as biomass molecules required for viability (S1 Table). We use the software package CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve all linear programming problems.

**Chemical environments and carbon sources**

To computationally predict the viability of a metabolism on a given carbon source, information about the chemical environment that contains this and other nutrients needed to synthesize biomass precursors is required. In our analysis of central carbon metabolism, we consider a minimal aerobic growth environment composed of a sole carbon source, along with ammonium as a nitrogen source, inorganic phosphate as a source of phosphorus, as well as oxygen, protons, and water. Different environments vary in their carbon source but are the same in other nutrients. A metabolism is viable on a given carbon source, if it can synthesize all 13 biomass precursors (Additional files 1 and 2) from that carbon source. In our study, we used the following carbon sources: D-glucose, acetate, pyruvate, D-lactate, D-fructose, alpha-ketoglutarate, fumarate, malate, succinate and glutamate.

**Reaction universe**

As a reaction "universe" we use a global set of reactions in central carbon metabolism, which is based on a published reconstruction of *E. coli* central carbon metabolism [29]. From this reconstruction, we deleted four reactions involved in ethanol metabolism, because in this study we are not interested in ethanol biosynthesis or degradation. We also grouped the reactions catalyzed by aconitase A and aconitase B into one reaction. The final reaction set consists of $N = 51$ intracellular reactions (Additional files 1 and 2). The reconstruction in [29] also involves 20 transport reactions, which are necessary to import nutrients or excrete waste products, and which we assume to be present in all metabolisms we study.

**Genotype-phenotype mapping in metabolic genotype space**

For computational expediency, we use a compact representation of a metabolic genotype that is based on reactions rather than genes. Specifically, we represent such a genotype as a binary vector whose *i-th* entry corresponds to the *i-th* reaction in our reaction universe. If an organism's genome encodes an enzyme capable of catalyzing a given reaction, the corresponding entry in the genotype vector will be one and zero otherwise. The genotype space including all possible metabolisms comprises $2^N$ metabolisms, where $N$ is the total number of known or considered chemical reactions ($N = 51$ for our analysis). Any metabolic genotype can be thought of as a point in this space. We consider a metabolism viable on a carbon source if its biomass synthesis rate is more than one percent of the biomass synthesis rate of the network formed by all $N$=51 reactions in central carbon metabolism. Some metabolic genotypes correspond to metabolisms in which some reactions are disconnected.

We call a reaction disconnected if (*i*) its products are neither biomass precursors nor substrates of any other reaction of the metabolism, or (*ii*) at least one of its substrates is neither a product of other reactions nor a nutrient taken up from the environment. We performed some analyses separately for metabolisms with and without disconnected reactions, to find out whether the presence of such reactions would affect our conclusions.

**Exhaustive enumeration of viable metabolisms**

To exhaustively characterize the phenotypes of all $2^{51}$ ($10^{15}$) metabolic genotypes would be prohibitive if one had to perform one FBA computation (consuming about $10^{-2}$ seconds) for each genotype. However, this computation becomes feasible when two facts are considered [39, 40]. First, six among the 51 internal reactions of central carbon metabolism are essential for viability on every carbon source we consider [62], which reduces the number of required FBA computations by a factor $2^6$ from $2^{51}$ ($10^{15}$) to $2^{45}$ ($10^{13}$). Second, deleting one or more reactions from an inviable metabolism cannot result in a viable metabolism, such that all metabolisms ("children") that contain a subset of the reactions of an inviable metabolism ("parent") will also be inviable. An algorithm that takes this observation into account decreases the number of required FBA evaluations further to approximately$10^9$ [41].

The set of metabolisms that are viable on a given subset *S* of the 10 carbon sources can easily be identified after the set $GN(C_i)$ of metabolisms viable on each of the 10 carbon sources $C_i$ has been determined. Specifically,
$V(S) = \{G \in \Omega, \forall\, C_i \in S, \forall\, C_j \in S' | G \in GN(C_i), G \notin GN(C_j)\}$ where *G* denotes a genotype from genotype space Ω, and *S'* denotes the complement of *S*.

## 7.6. References


1. Wagner A: *Arrival of the Fittest: Solving Evolution's Greatest Puzzle*. 1st edition. London: Oneworld Publications; 2014.

2. Gould SJ, Vrba ES: Exaptation; a missing term in the science of form. *Paleobiology* 1982, 8:4–15.

3. Bock WJ: Preadaptation and Multiple Evolutionary Pathways. *Evolution (N Y)* 1959, 13(June):194–211.

4. Darwin C: *Darwin Online: On the Origin of Species*. 6th edition. London: Adamant Media Corporation; 1872.

5. True JR, Carroll SB: Gene co-option in physiological and morphological evolution. *Annu Rev Cell Dev Biol* 2002, 18:53–80.

6. Zákány J, Duboule D: Hox genes in digit development and evolution. *Cell Tissue Res* 1999, 296:19–25.

7. Keys DN, Lewis DL, Selegue JE, Pearson BJ, Goodrich L V, Johnson RL, Gates J, Scott MP, Carroll SB: Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* 1999, 283:532–4.

8. Tomarev SI, Piatigorsky J: Lens crystallins of invertebrates--diversity and recruitment from detoxification enzymes and novel proteins. *Eur J Biochem* 1996, 235:449–65.

9. Pievani T, Serrelli E: Exaptation in human evolution: how to test adaptive vs exaptive evolutionary hypotheses. *J Anthropol Sci* 2011, 89:9–23.

10. Schuster P, Fontana W, Stadler PF, Hofacker IL: From sequences to shapes and back: a case study in RNA secondary structures. *Proc Biol Sci* 1994, 255:279–84.

11. Lipman DJ, Wilbur WJ: Modelling neutral and selective evolution of protein folding. *Proc Biol Sci* 1991, 245:7–11.

12. Cowperthwaite MC, Economo EP, Harcombe WR, Miller EL, Meyers LA: The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comput Biol* 2008, 4:e1000110.

13. Ferrada E, Wagner A: A comparison of genotype-phenotype maps for RNA and proteins. *Biophys J* 2012, 102:1916–25.

14. Samal A, Matias Rodrigues JF, Jost J, Martin OC, Wagner A: Genotype networks in metabolic reaction spaces. *BMC Syst Biol* 2010, 4:30.

15. Copley SD: Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem Sci* 2000, 25:261–5.

16. Rehmann L, Daugulis AJ: Enhancement of PCB degradation by Burkholderia xenovorans LB400 in biphasic systems by manipulating culture conditions. *Biotechnol Bioeng* 2008, 99:521–8.

17. Van der Meer JR, Werlen C, Nishino S, Spain J: Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. *Appl Environ Microbiol* 1998, 64:4185–93.

18. Cline RE, Hill RH, Phillips DL, Needham LL: Pentachlorophenol measurements in body fluids of people in log homes and workplaces. *Arch Environ Contam Toxicol* , 18:475–81.

19. Detkova EN, Boltyanskaya Y V.: Osmoadaptation of haloalkaliphilic bacteria: Role of osmoregulators and their possible practical application. *Microbiology* 2007, 76:511–522.

20. Dantas G, Sommer MOA, Oluwasegun RD, Church GM: Bacteria subsisting on antibiotics. *Science* 2008, 320:100–3.

21. Feist AM, Palsson BØ: The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* 2008, 26:659–67.

22. Oberhardt MA, Palsson BØ, Papin JA: Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009, 5:320.

23. McCloskey D, Palsson BØ, Feist AM: Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol Syst Biol* 2013, 9:661.

24. Lewis NE, Nagarajan H, Palsson BO: Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol* 2012, 10:291–305.

25. Wagner A: Metabolic networks and their evolution. *Adv Exp Med Biol* 2012, 751:29–52.

26. Matias Rodrigues JF, Wagner A: Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput Biol* 2009, 5:e1000613.

27. Edwards JS, Palsson BO: The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A* 2000, 97:5528–33.

28. Edwards JS, Palsson BO: Systems properties of the Haemophilus influenzae Rd metabolic genotype. *J Biol Chem* 1999, 274:17410–6.

29. Orth JD, Fleming RMT, Palsson BØ: Reconstruction and Use of Microbial Metabolic Networks: the Core Escherichia coli Metabolic Model as an Educational Guide. *EcoSal Plus* 2010.

30. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ: A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, 3:121.

31. Edwards JS, Ibarra RU, Palsson BO: In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001, 19:125–30.

32. Segrè D, Vitkup D, Church GM: Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 2002, 99:15112–7.

33. Förster J, Famili I, Fu P, Palsson BØ, Nielsen J: Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res* 2003, 13:244–53.

34. Wang Z, Zhang J: Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol Evol* 2009, 1:23–33.

35. Papp B, Pál C, Hurst LD: Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 2004, 429:661–4.

36. Barve A, Wagner A: A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* 2013, 500:203–6.

37. Papagianni M: Recent advances in engineering the central carbon metabolism of industrially important bacteria. *Microb Cell Fact* 2012, 11:50.

38. Noor E, Eden E, Milo R, Alon U: Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol Cell* 2010, 39:809–20.

39. Barve A, Hosseini S-R, Martin OC, Wagner A: Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. *BMC Syst Biol* 2014, 8:48.

40. Hosseini S-R, Barve A, Wagner A: Exhaustive Analysis of a Genotype Space Comprising 1015 Central Carbon Metabolisms Reveals an Organization Conducive to Metabolic Innovation. *PLoS Comput Biol* 2015, 11:e1004329.

41. Hosseini S-R: Exhaustive genotype-phenotype mapping in metabolic genotype space. Swiss Federal Instiute of Technology; 2013.

42. Huynen MA, Dandekar T, Bork P: Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol* 1999, 7:281–91.

43. Sokal RR, Michener CD: A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Scientific Bulletin* 1958, 38:1409–1438.

44. Matias Rodrigues JF, Wagner A: Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst Biol* 2011, 5:39.

45. Wagner A, Andriasyan V, Barve A: The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism. *Journal of Molecular Biochemistry* 2014.

46. Ibarra RU, Edwards JS, Palsson BO: Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 2002, 420:186–9.

47. Fong SS, Palsson BØ: Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet* 2004, 36:1056–8.

48. Fong SS, Marciniak JY, Palsson BO: Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J Bacteriol* 2003, 185:6400–6408.

49. HUNTER RL, MARKERT CL: Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science* 1957, 125:1294–5.

50. Khersonsky O, Tawfik DS: Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu Rev Biochem* 2010, 79:471–505.

51. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD: Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis. *Mol Syst Biol* 2010, 6:436.

52. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, Palsson BO: Network context and selection in the evolution to enzyme specificity. *Science* 2012, 337:1101–4.

53. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, Lázár V, Spohn R, Csörgő B, Wagner A, Ruppin E, Pál C, Papp B: Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A* 2014, 111:11762–7.

54. Meléndez-Hevia E, Waddell TG, Heinrich R, Montero F: Theoretical approaches to the evolutionary optimization of glycolysis--chemical analysis. *Eur J Biochem* 1997, 244:527–43.

55. Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R: Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc Natl Acad Sci U S A* 2013, 110:10039–44.

56. Ebenhöh O, Handorf T: Functional classification of genome-scale metabolic networks. *EURASIP J Bioinform Syst Biol* 2009, 2009:570456.

57. Reed JL, Vo TD, Schilling CH, Palsson BO, Palsson B, Palsson B, Edwards J, Covert M, Palsson B, Varma A, Palsson B, Bonarius H, Schmid G, Tramper J, Price N, Papin J, Schilling C, Palsson B, Reed J, Palsson B, Edwards J, Palsson B, Edwards J, Ibarra R, Palsson B, Ibarra R, Edwards J, Palsson B, Serres M, Gopal S, et al.: An expanded genome-scale model of Escherichia coli K-12 ( i JR904 GSM/GPR). *Genome Biol* 2003, 4:R54.

58. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* 2007, 2:727–38.

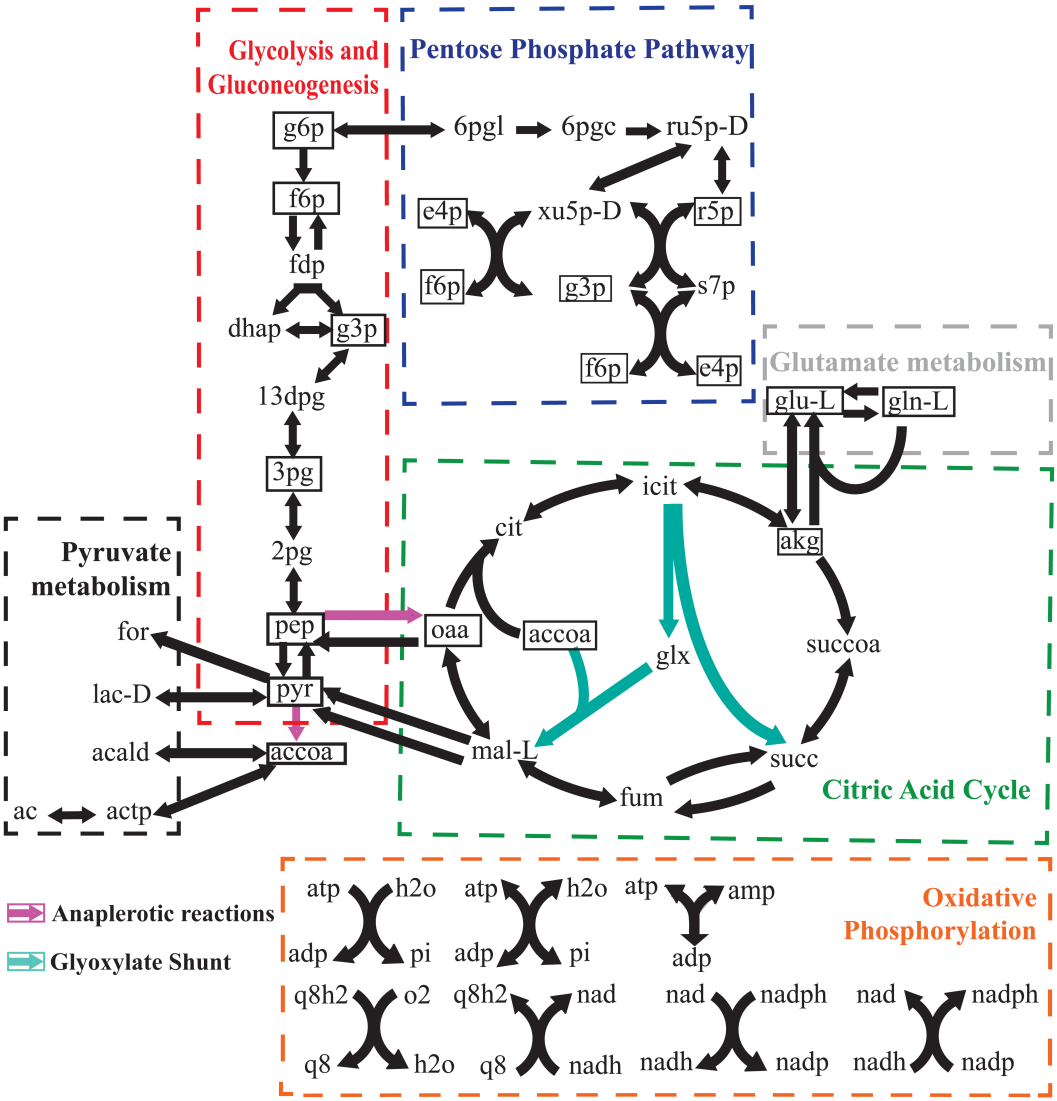
59. Price ND, Reed JL, Palsson BØ: Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004, 2:886–97.

60. Edwards JS, Covert M, Palsson B: Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol* 2002, 4:133–40.

61. Orth JD, Thiele I, Palsson BØ: What is flux balance analysis? *Nat Biotechnol* 2010, 28:245–8.

62. Barve A, Rodrigues JFM, Wagner A: Superessential reactions in metabolic networks. *Proc Natl Acad Sci U S A* 2012, 109:E1121–30.

## 7.7. Supplementary Information

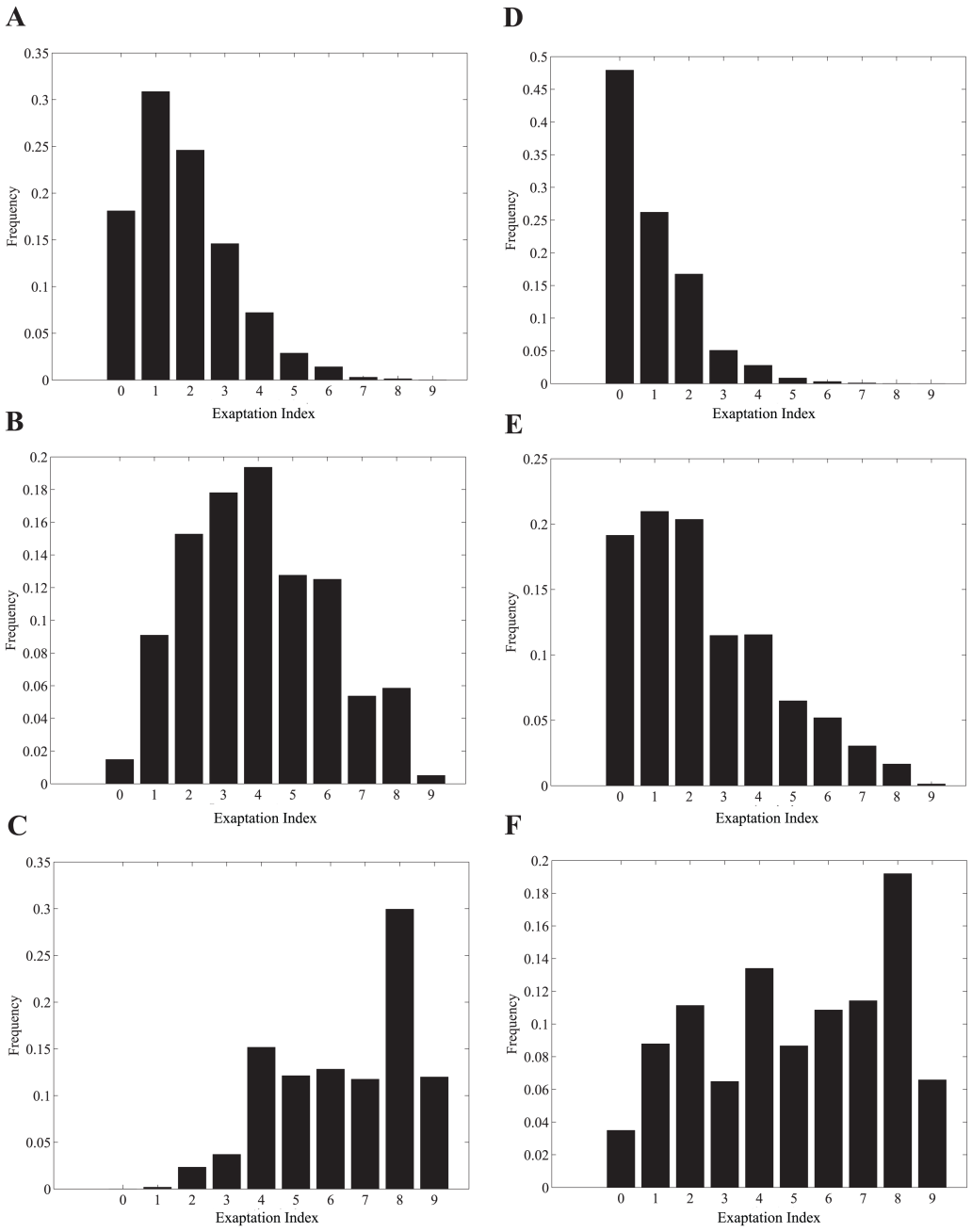



**Additional file 1: Central carbon metabolism.** Each arrow in each panel corresponds to one of the 51 internal reactions we consider. Metabolites are indicated by their acronyms (see Additional file 2). Boxed metabolites correspond to 13 essential biomass precursors. Note that 4 metabolites (accoa, g3p, f6p and e4p) are shown more than once for visual clarity. Metabolic pathways, including glycolysis/gluconeogenesis, pentose-phosphate pathway, citric-acid cycle, oxidative phosphorylation, pyruvate and glutamate metabolism are distinguished by the colored and dashed rectangles. Anaplerotic reactions and glyoxylate shunt are highlighted using the purple and green arrows respectively. The figure is taken by permission from [40].

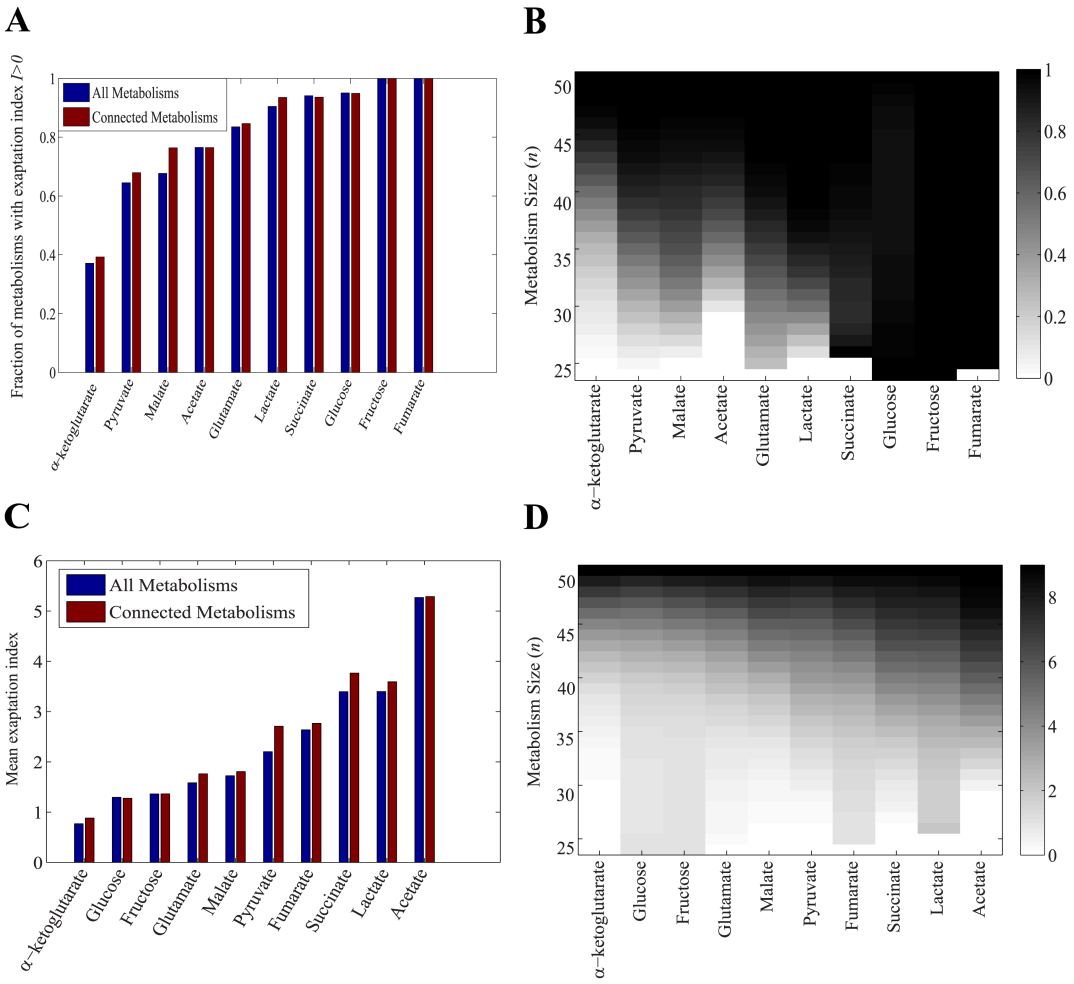
| Abbreviation | Metabolite full name |
|---|---|
| 13dpg | 3-Phospho-D-glyceroyl phosphate |
| 2pg | D-Glycerate 2-phosphate |
| 3pg | 3-Phospho-D-glycerate |
| 6pgc | 6-Phospho-D-gluconate |
| 6pgl | 6-phospho-D-glucono-1,5-lactone |
| ac | Acetate |
| ac[e] | Acetate (extracellular) |
| acald | Acetaldehyde |
| acald[e] | Acetaldehyde (extracellular) |
| accoa | Acetyl-CoA |
| actp | Acetyl phosphate |
| adp | ADP |
| akg | 2-Oxoglutarate |
| akg[e] | 2-Oxoglutarate (extracellular) |
| amp | AMP |
| atp | ATP |
| cit | Citrate |
| co2 | CO2 |
| co2[e] | CO2 (extracellular) |
| coa | Coenzyme A |
| dhap | Dihydroxyacetone phosphate |
| e4p | D-Erythrose 4-phosphate |
| f6p | D-Fructose 6-phosphate |
| fdp | D-Fructose 1,6-bisphosphate |
| for | Formate |
| for[e] | Formate (extracellular) |
| fru[e] | D-Fructose (extracellular) |
| fum | Fumarate |
| fum[e] | Fumarate (extracellular) |
| g3p | Glyceraldehyde 3-phosphate |
| g6p | D-Glucose 6-phosphate |
| glc-D[e] | D-Glucose (extracellular) |
| gln-L | L-Glutamine |
| gln-L[e] | L-Glutamine (extracellular) |
| glu-L | L-Glutamate |
| glu-L[e] | L-Glutamate (extracellular) |
| glu-L[e] | L-Glutamate (extracellular) |
| glx | Glyoxylate |
| h | H+ |
| h[e] | H+ (extracellular) |
| h2o | H2O |
| h2o[e] | H2O (extracellular) |
| icit | Isocitrate |
| lac-D | D-Lactate |
| lac-D[e] | D-Lactate (extracellular) |
| mal-L | L-Malate |
| mal-L[e] | L-Malate (extracellular) |
| nad | Nicotinamide adenine dinucleotide |
| nadh | Nicotinamide adenine dinucleotide (reduced) |
| nadp | Nicotinamide adenine dinucleotide phosphate |
| nadph | Nicotinamide adenine dinucleotide phosphate (reduced) |
| nh4 | Ammonium |
| nh4[e] | Ammonium (extracellular) |
| o2 | O2 |
| o2[e] | O2 (extracellular) |
| oaa | Oxaloacetate |

| | |
|---|---|
| pep | Phosphoenolpyruvate |
| pi | Phosphate |
| pi[e] | Phosphate (extracellular) |
| pyr | Pyruvate |
| pyr[e] | Pyruvate (extracellular) |
| q8 | Ubiquinone-8 |
| q8h2 | Ubiquinol-8 |
| r5p | alpha-D-Ribose 5-phosphate |
| ru5p-D | D-Ribulose 5-phosphate |
| s7p | Sedoheptulose 7-phosphate |
| succ | Succinate |
| succ[e] | Succinate (extracellular) |
| succoa | Succinyl-CoA |
| xu5p-D | D-Xylulose 5-phosphate |

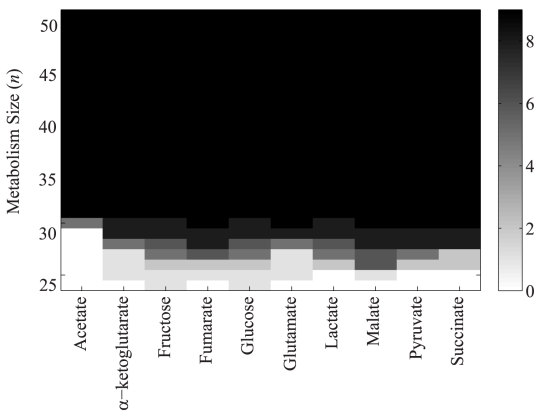
**Additional file 2: Metabolites and reactions in central carbon metabolism.** This table lists all the reactions (Sheet 1) and metabolites (Sheet 2) of central carbon metabolism, their abbreviations, and further associated information.



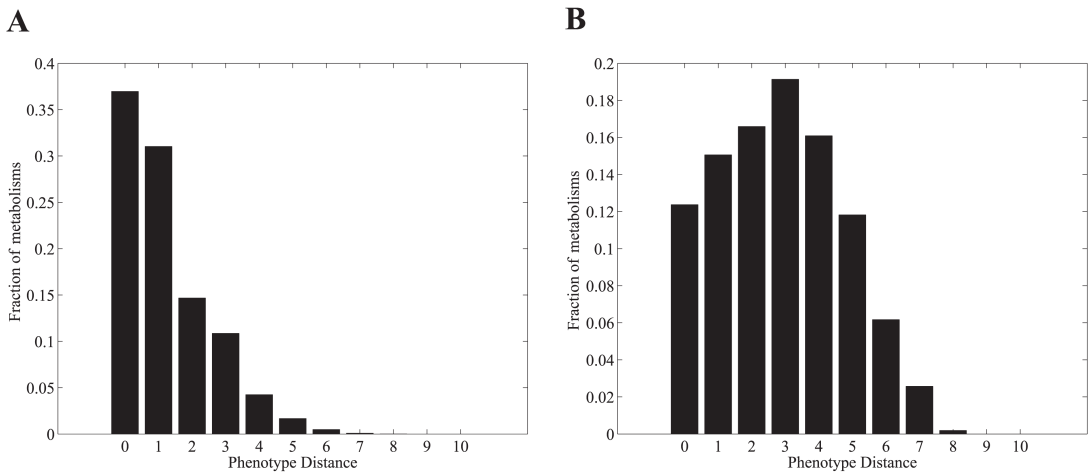

**Additional file 3: Metabolisms viable on a given carbon source vary widely in their exaptation potential.** Histogram of the exaptation index (*x*-axis), i.e., the number of carbon sources $C_{new}$ on which a metabolism is viable, for metabolisms viable on lactate as carbon source *C* with size **(A)** 35, **(B)** 40, **(C)** 45, and for metabolisms viable on malate as carbon source *C* with size **(D)** 35, **(E)** 40, **(F)** 45.



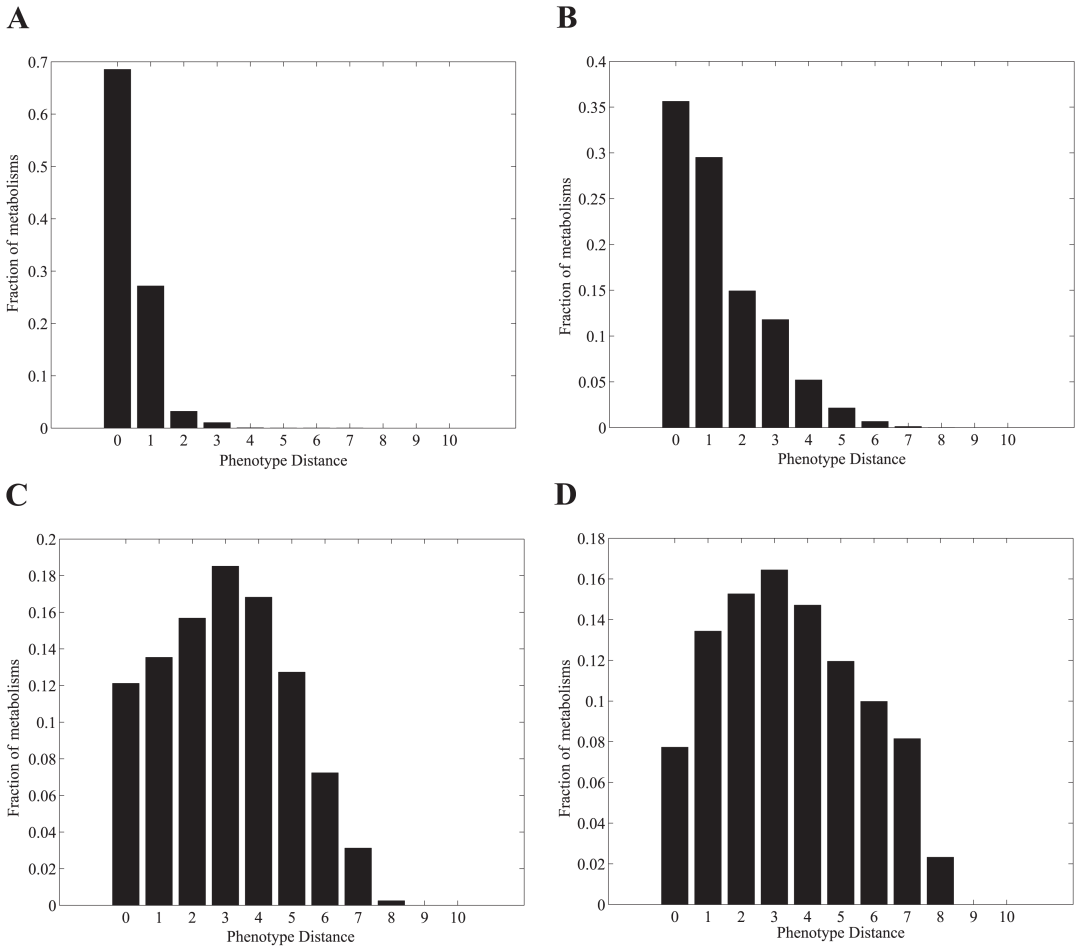

**Additional file 4: High exaptation potential in central carbon metabolisms (considering only metabolisms without disconnected reactions). (A)** Fraction of metabolisms with exaptation index $I>0$ ($y$-axis) viable on some carbon source $C$ ($x$-axis). Red bars correspond to viable metabolisms without disconnected reactions, and blue bars correspond to all viable metabolisms. **(B)** Fraction of metabolisms (coded by shade of grey, see legend) with exaptation index ($I>0$) that are viable on some carbon source $C$ ($x$-axis) and have a given size ($y$-axis), . **(C)** Mean exaptation index of metabolisms without disconnected reactions, and viable on a given focal carbon source $C$ ($x$-axis). Red bars correspond to viable metabolisms without disconnected reactions, and blue bars correspond to all viable metabolisms. **(D)** Mean exaptation index (coded by shade of grey, see legend) of metabolisms without disconnected reactions, and viable on some carbon source $C$ ($x$-axis) and with a given size ($y$-axis). White colors in (B) and (D) correspond to metabolisms whose size is too small for viability on $C$.



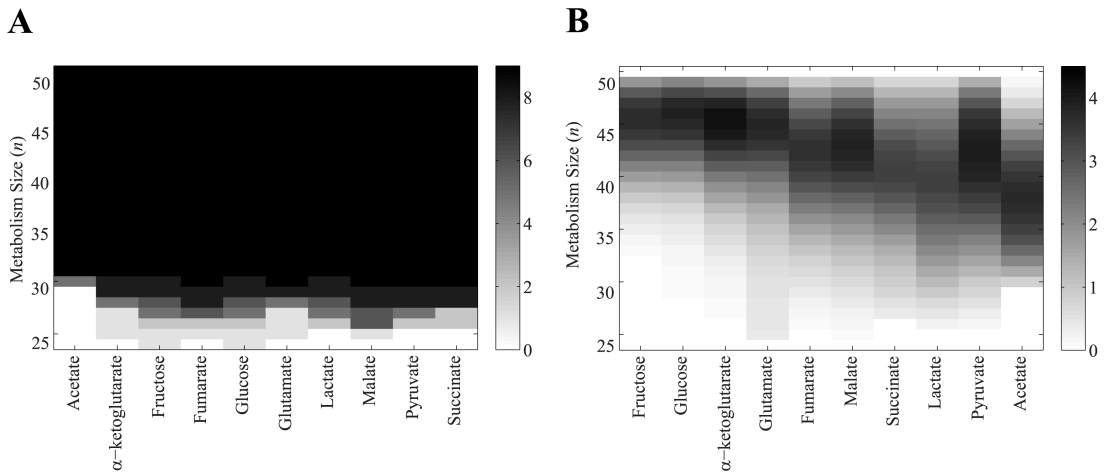

**Additional file 5: Exaptation diversity.** For metabolisms whose focal carbon source *C* is shown on the *x*-axis, and the number of reactions (*n*) is shown on the vertical axis, the number of carbon sources $C_{new}$ on which at least one metabolism is preadapted, is coded by shade of grey (see legend).



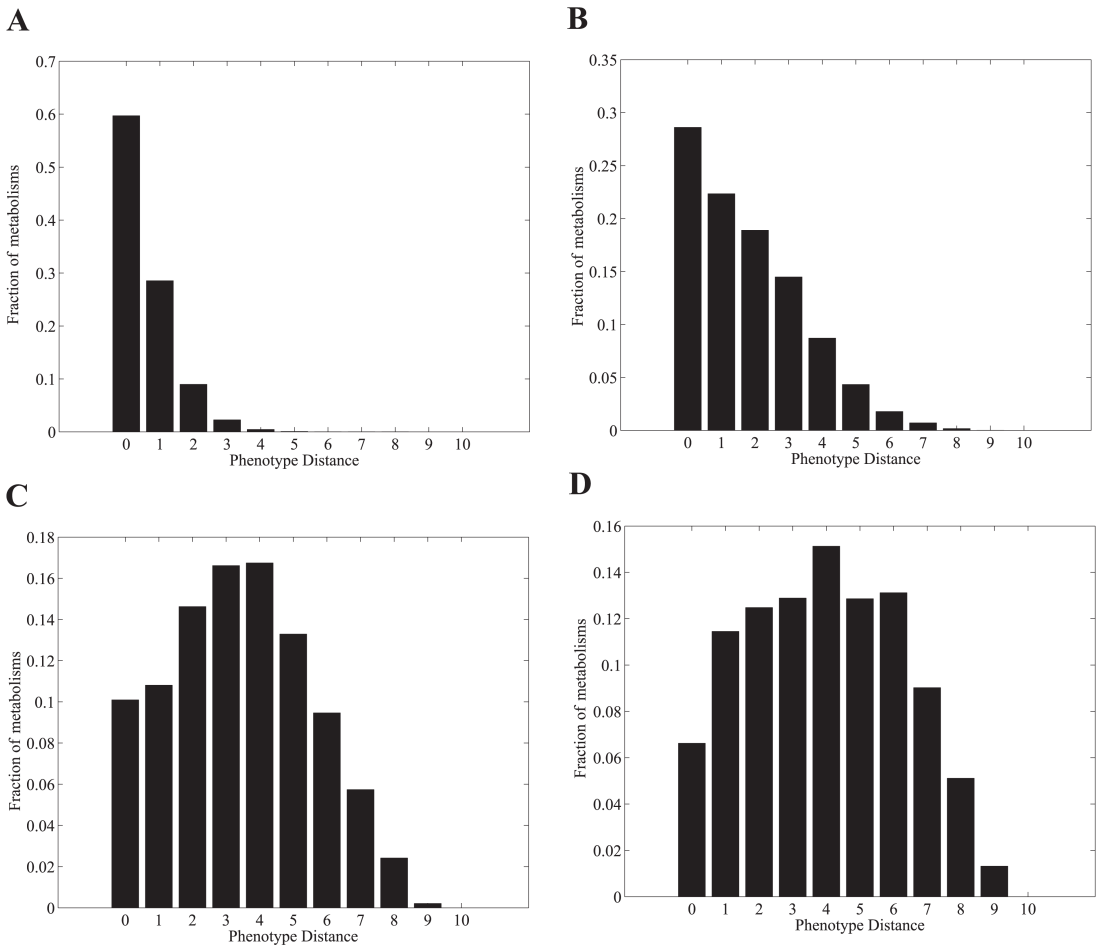

**Additional file 6**: **Metabolisms viable on glucose as the main carbon source *C* can differ greatly in their viability on other carbon sources.** The figure shows a histogram of the phenotype distance (*x*-axis), for metabolisms of size **(A)** 35, and **(B)** 40, viable on glucose as carbon source *C*.



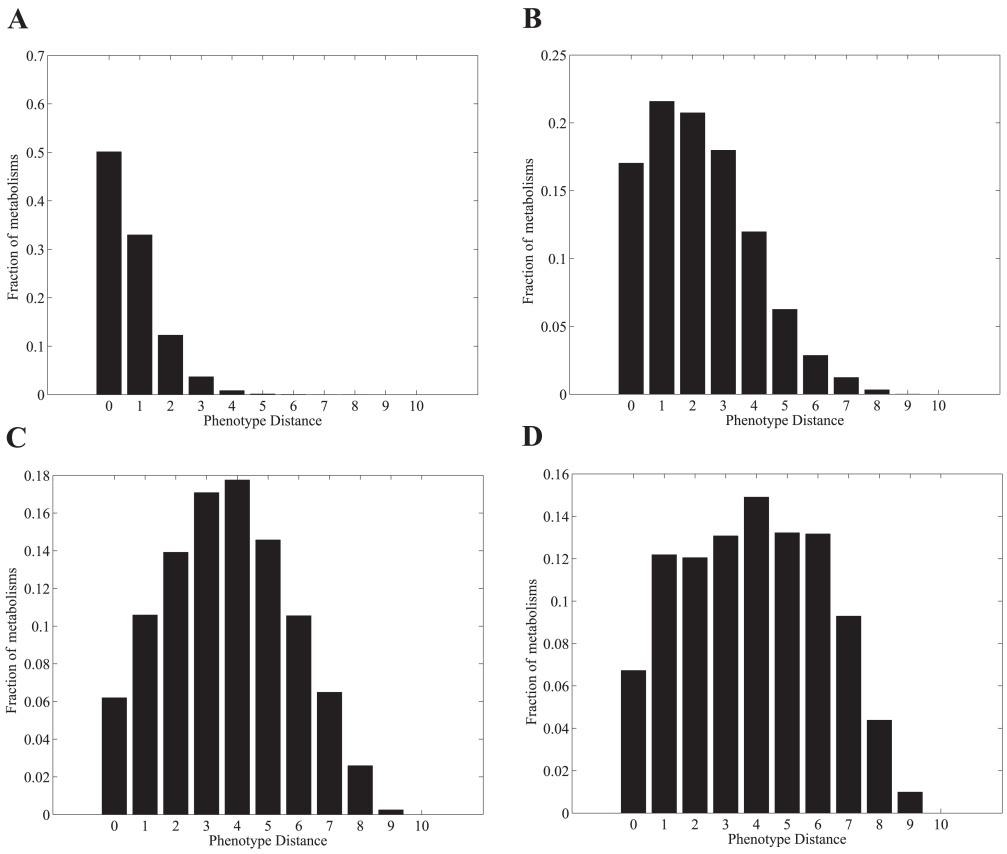

**Additional file 7**: **Metabolisms viable on pyruvate as the main carbon source *C* can differ greatly in their viability on other carbon sources.** The figure shows a histogram of the phenotype distance (*x*-axis), for metabolisms of size **(A)** 30, **(B)** 35, **(C)** 40, and **(D)** 45, viable on pyruvate as focal carbon source C.



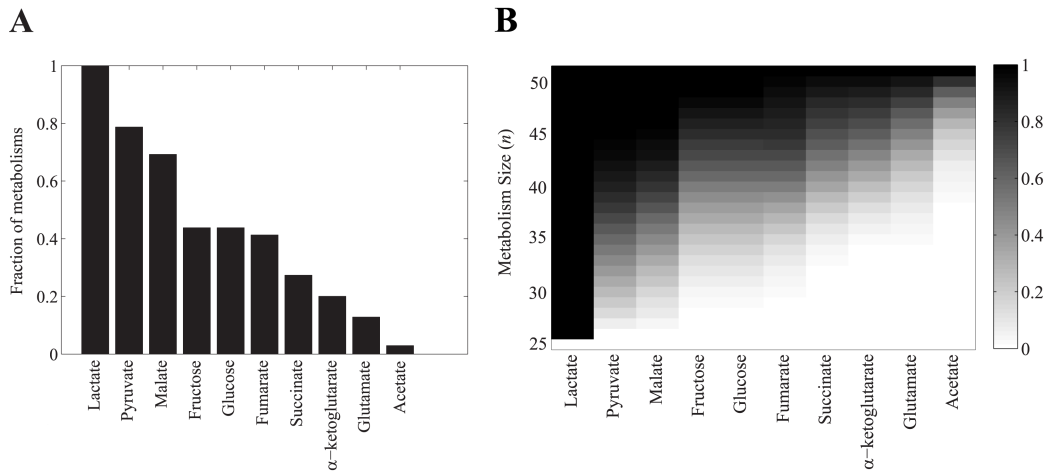

**Additional file 8: Metabolisms can preadapt to a wide variety of carbon sources (considering only metabolisms without disconnected reactions). (A)** For metabolisms (without disconnected reactions) whose focal carbon source $C$ is shown on the $x$-axis, and the number of reactions ($n$) is shown on the vertical axis, each shade of grey (see legend) shows the number of carbon sources $C_{new}$ on which at least one metabolism is preadapted. **(B)** Mean phenotypic distance (see legend) of metabolisms (without disconnected reactions) viable on a focal carbon source ($x$-axis) and with a given number of reactions $n$ ($y$-axis).



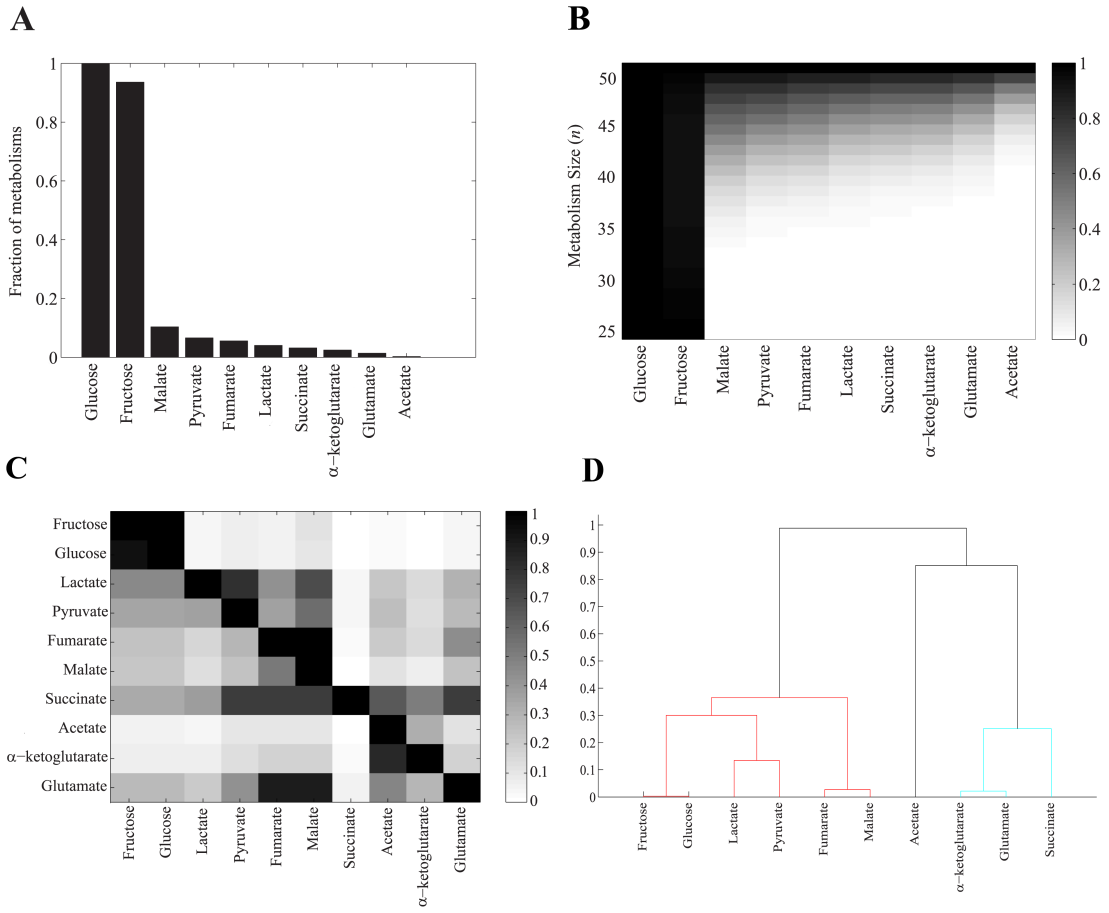

**Additional file 9**: **Metabolisms viable on glucose as the main carbon source *C* can differ greatly in their viability on other carbon sources (considering only metabolisms without disconnected reactions).** The figure shows a histogram of the phenotype distance (*x*-axis), for metabolisms without disconnected reactions with size **(A)** 30, **(B)** 35, **(C)** 40, and **(D)** 45, viable on glucose as carbon source *C*.



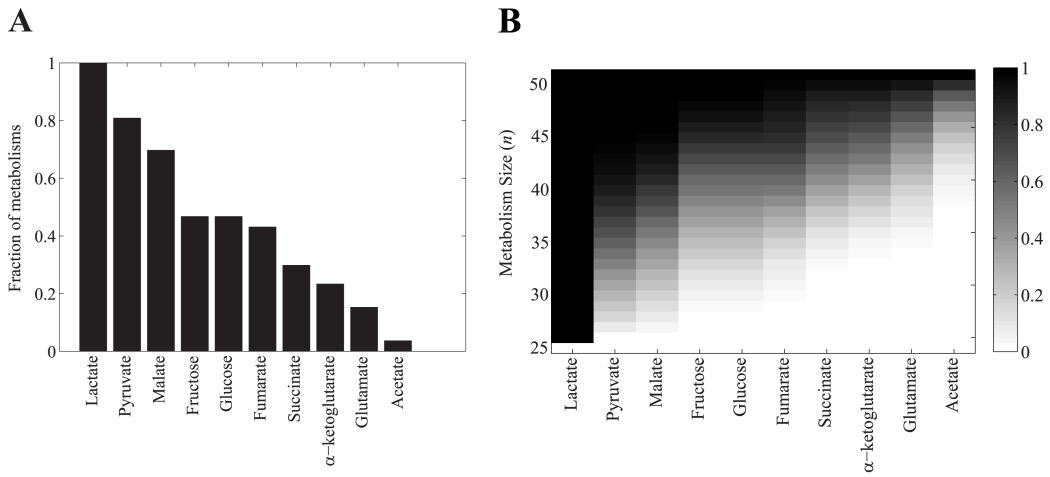

**Additional file 10**: **Metabolisms viable on pyruvate as the main carbon source *C* can differ greatly in their viability on other carbon sources (considering only metabolisms without disconnected reactions).** The figure shows a histogram of the phenotype distance (*x*-axis), for metabolisms without disconnected reactions with size **(A)** 30, **(B)** 35, **(C)** 40, and **(D)** 45, viable on pyruvate as carbon source *C*.



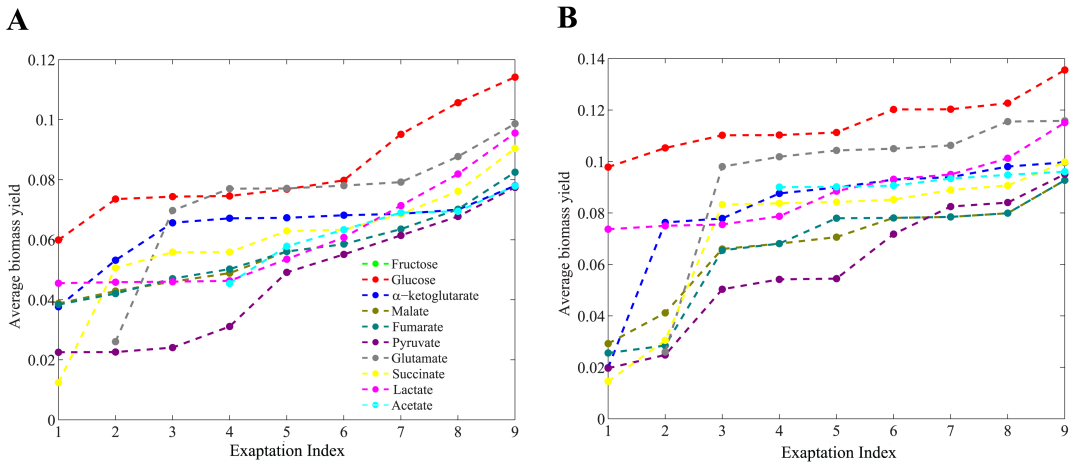

**Additional file 11: Metabolisms viable on lactate differ in their propensity for preadaptation to other carbon sources $C_{new}$. (A)** The histogram shows the fraction of metabolisms viable on lactate as carbon source *C* that are also viable on each of the nine other carbon sources $C_{new}$ (*x*-axis). **(B)** As in (A), but broken down by metabolism size, and fractions of viable metabolisms are coded by shade of grey, see legend.



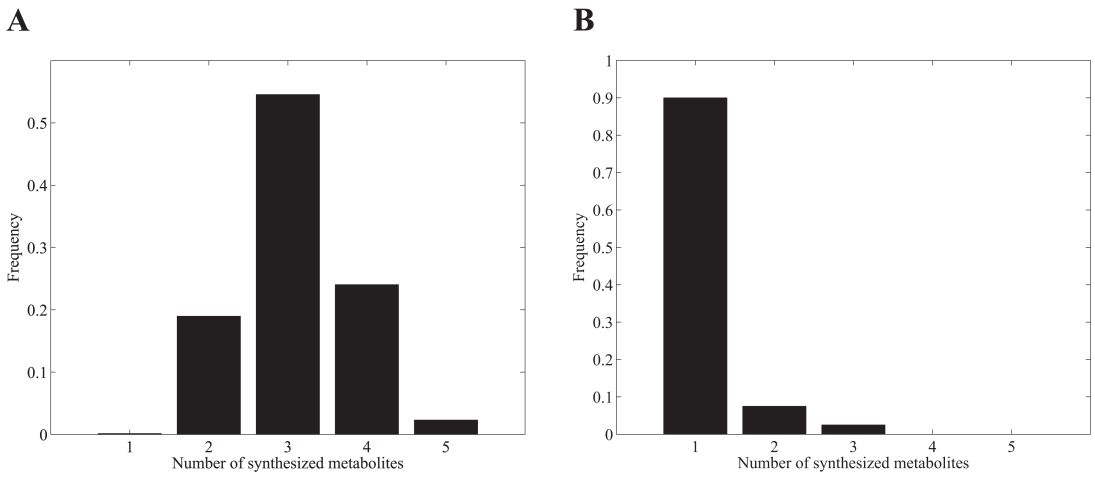

**Additional file 12: Potential for preadaptation depends on biochemical similarity between carbon sources (considering only metabolisms without disconnected reactions). (A)** The histogram shows the fraction of metabolisms (without disconnected reactions) viable on glucose as carbon source $C$ that are also viable on each of the nine other carbon sources $C_{new}$ ($x$-axis). **(B)** As in (A), but broken down by metabolism size, and fractions of viable metabolisms are coded by shade of grey, see legend **(C)** Fraction of metabolisms (without disconnected reactions) viable on carbon source $C$ ($x$-axis), that are also viable on carbon source $C_{new}$ ($y$-axis), are coded by shade of grey, see legend. **(D)** Dendrogram of carbon sources clustered based on their pairwise preadaptation propensity. We used UPGMA method (unweighted pair group method with arithmetic means), for clustering carbon sources.



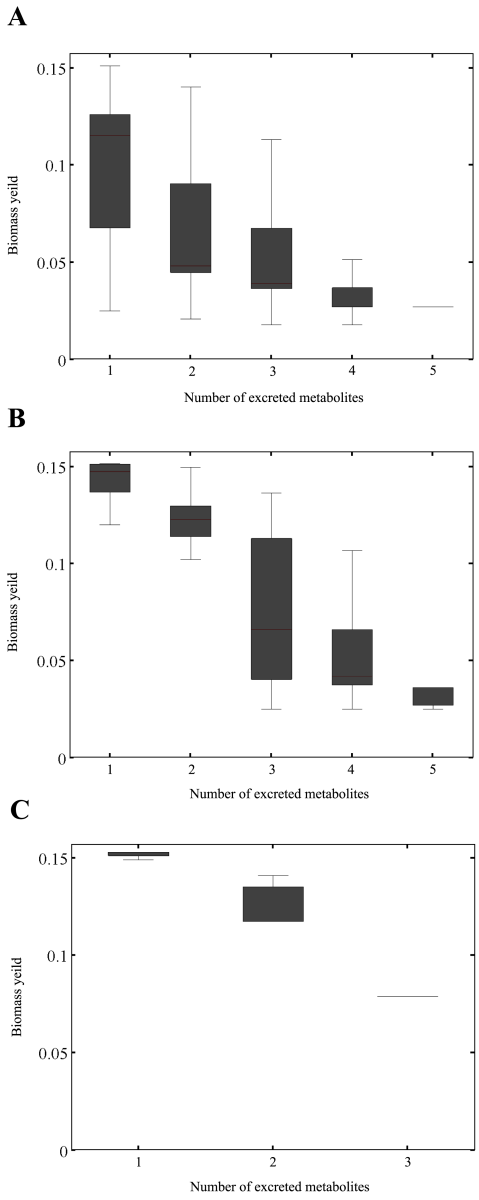

**Additional file 13: Metabolisms viable on lactate have different potential for preadaptation to other carbon sources $C_{new}$ (considering only metabolisms without disconnected reactions). (A)** The histogram shows the fraction of metabolisms (without disconnected reactions) viable on lactate as carbon source $C$ that are also viable on each of the nine other carbon sources $C_{new}$ ($x$-axis). **(B)** As in (A), but broken down by metabolism size, and fractions of viable metabolisms are coded by shade of grey, see legend.



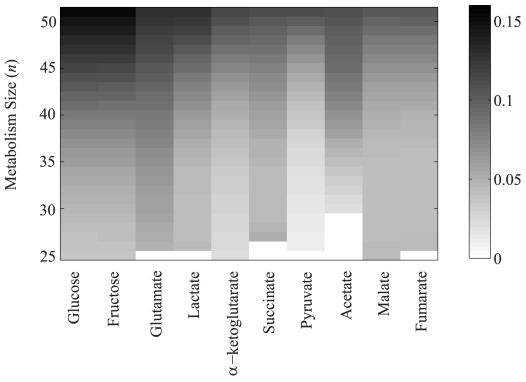

**Additional file 14: Association between exaptation potential and biomass yield.** The $x$-axes show the exaptation index, i.e., the number of carbon sources $C_{new}$ on which metabolisms viable on carbon source $C$ (color legend) are viable. The $y$-axes show the average biomass yield. Data is based on metabolisms of size **(A)** $n=35$, and **(B)** $n=45$.




**Additional file 15: Distribution of the number of synthesized metabolites among metabolisms viable on glucose.** Fraction of metabolisms excreting a given number of metabolites (*x*-axis) among metabolisms viable on glucose with size **(A)** *n=35*, and **(B)** *n=45*.




**Additional file 16: Association between biomass yield and number of excreted waste metabolites.** The vertical axis shows the biomass yield of metabolisms of size **(A)** *n=30*, **(B)** *n=40*, and **(C)** *n=50* viable on glucose that excrete a given number of metabolites (*x*-axis). Boxes span the *25*-th to *75*-th percentile, and whiskers indicate the maximu shows the m and minimum values.




**Additional file 17: Larger metabolisms have higher biomass yield.** Mean of biomass yield among metabolisms of a given size (*y*-axis), that are viable on a given carbon source (*x*-axis), coded by shade of grey, see legend. White colors correspond to metabolisms whose size is too small for viability on *C*.

# Chapter 8: Conclusion

In conclusion, in my thesis, I have studied the origins of evolutionary innovations in complex metabolic systems. In particular, my studies have identified the prominent principles underlying the emergence of novel metabolic phenotypes. First and foremost, I have shown that recombination is a powerful mechanism of genetic change behind phenotypic innovation. My results revealed that recombination is able to dramatically increase the probability of the emergence of phenotypically innovative recombinant genotypes. This ability reflects the combinatorial nature of innovation in biological systems, that is, novel traits emerge when already-existing components come together in a new combination. Moreover, my results identify the factors that can enhance phenotypic innovation through recombination. These principles can equip us with the capability to manipulate and control biological systems towards higher innovation capacity. For example, my results revealed that recombination between genotypically more similar but phenotypically more dissimilar metabolisms can lead to a higher probability of the emergence of novel phenotypes. These observations may find practical value in experimental settings to generate pools of recobinant genomes with the capacity to manifest novel phenotypes.

Moreover, my exhaustive analysis of genotype space in central carbon metabolism revealed organizational principles facilitating metabolic innovations. In particular, metabolic genotypes are organized in genotype space in a non-random way, which ensures high probability of the emergence of novel phenotypes. My observations indicate that genotypes with the same phenotype form a connected network in genotype space. In other words, the members of every pair of viable genotypes are accessible from each other through a few viability-preserving mutational steps. This property renders the emergence of novel metabolic phenotypes independent of historical events, and so historical contingency does not play a strong role in the emergenece of metabolic properties. Moreover, all possible metabolic phenotypes are accessible in the immediate neighborhood of most viable genotypes. Thus, novel phenotypes are accessible in genotype space through a small number of genetic changes.

Furthermore, I observed that metabolic traits are highly correlated, and this correlation stems from overlapping network components, for example shared metabolic reactions and pathways that connect biomass components to external metabolites. The extent of this correlation raises the possibility that metabolic phenotypes might have non-adaptive origins. In other words, the emergence of novel phenotypes in metabolic systems may not necessarily have adaptive reasons. Instead, it can happen passively due to the inherent correlation between metabolic phenotypes.

Finally, my results revealed that in metabolic systems, phenotypic robustness helps promote phenotypic innovation. Importantly, genomes have evolved an organization that ensures substantially higher phenotypic robustness to large-scale gene deletions than a random arrangement of metabolic genes. This observation provides evidence supporting the claim that randomness is not sufficient to explain complex phenotypes in biological systems. An evolved arrangement of metabolic genes provides higher phenotypic robustness to the deleterious effects of large-scale gene deletions, and this increased robustness can be helpful for the emergence of novel phenotypes.

In sum, the take home message of my dissertation is that a complex biological phenomenon like metabolic innovation can be analyzed rigorously in order to dissect underlying evolutionary rules and principles. If we characterize the hidden rules behind it, phenotypic innovation will no longer remain a black box. Although evolutionary theory is the ultimate explanation for the emergence of complex biological traits, we still need to identify the proximate cuases behind phenotypic innovation. Importantly, characterizing the proximate causes of innovation can mechanistically clarify the underlying processes, and provides us with a more accurate and evidence-based interpretation of evolutionary theory in the context of complex systems.

# Acknowledgements

I am grateful to all my teachers, mentors and supervisors, most prominently to my Ph.D. supervisor, Prof. Andreas Wagner. I have gained fantastic research experiences in his lab, from technically demanding projects to conceptually challenging problems and theory development. I particularly thank him for his efforts in improving my scientific writing.

Also, I would like to thank my Ph.D. committee members, Prof. Olivier Martin and Prof. Frédéric Guillaume for their useful comments and suggestions and for their flexibility and robustness to changes in my defense date!

I am thankful to Ms. Annette Schmidt for her kind attitude towards my administrative issues, and to IT management team in Wagner group, Markus Neumann and Vinzenz Muser for their unconditional assistance and support.

I wish to thank all my friends and colleagues in Irchel Campus. In particular, Michael Hediger for the statistical discussions and for the German translations, my officemates, Jia Zheng and Pierre Laye and all other members of Wagner lab for all their support during the four years of my PhD studies at the University of Zurich.

# Curriculum Vitae

**First Name:** Sayed-Rzgar
**Last Name**: Hosseini
**Date of Birth:** 18/12/1986
**Place of Birth:** Saghez, Kurdistan
**E-mail:** rzgar.hosseini@ieu.uzh.ch
**Researcher ID (ORCID ID):** 0000-0002-2308-6754

## Education:

- **09/2013-** Ph.D., Evolutionary Biology, University of Zurich
- **09/2015-** M.Sc., Statistics, ETH Zurich (60 ECTS passed; Master thesis remains)
- **09/2013-08/2014** CAS, Computer Science, ETH Zurich
- **09/2010-08/2013** M.Sc., Computational Biology and Bioinformatics, ETH Zurich
- **09/2005-07/2010** B.Sc. & M.Sc., Biotechnology, University of Tehran

## Employment History:

- **09/2013-08/2017** Institute of Evolutionary Biology and Environmental Studies (IEU), University of Zurich. Advisor: Prof. Dr. Andreas Wagner

## Teaching assistantships:

- Introduction to Bioinformatics I (ETH: 551-1295-00L, UZH: BCH401), Fall semesters 2014, 2015 and 2016.
- Practical Bioinformatics (UZH: Bio334), spring semesters 2015 and 2016.

## Publication:

### In preparation:

- **Hosseini, S.-R.,** Payne Jl, and A. Wagner. Exhaustive fitness landscape analysis in central carbon metabolism.
- **Hosseini, S.-R.,** and A. Wagner. Robustness and evolvability in central carbon metabolism.
- **Hosseini, S.-R.,** and A. Wagner. An efficient algorithm for systematic construction of minimal bacterial metabolic genomes.
- Libby E, Hubert-Dufresne L, **Hosseini, S.-R.,** and A. Wagner. Syntrophy emerges spontaneously in complex metabolic systems.

### In Revision:

- **Hosseini, S.-R.,** and A. Wagner. Genomic organization underlying deletional robustness in bacterial metabolic systems. Under review in PNAS.

**Published:**

- **Hosseini, S.-R.,** and A. Wagner. 2017. Constraint and Contingency Pervade the Emergence of Novel Phenotypes in Complex Metabolic Systems. Biophys. J. 113: 690–701. http://www.cell.com/biophysj/fulltext/S0006-3495(17)30684-7
- **Hosseini, S.-R.,** O. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc. R. Soc. B. http://rspb.royalsocietypublishing.org/content/283/1839/20161536.long
- **Hosseini, S.-R.,** and A. Wagner. 2016. The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. BMC Syst. Biol. 10: 97. https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-016-0343-7
- **Hosseini, S.-R.,** A. Barve, and A. Wagner. 2015. Exhaustive analysis of a genotype space comprising $10^{15}$ central carbon metabolisms reveals an organization conducive to metabolic innovation. PLoS Comput. Biol. 11: e1004329. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004329
- Barve, A., **S.-R. Hosseini**, O.C. Martin, and A. Wagner. 2014. Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. BMC Syst. Biol. 8: 48. https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-48
- Ghazavi, F., Z. Fazlali, S.S. Banihosseini**, S.-R. Hosseini,** M.H. Kazemi, S. Shojaee, K. Parsa, H. Sadeghi, F. Sina, M. Rohani, G.-A. Shahidi, N. Ghaemi, M. Ronaghi, and E. Elahi. 2011. PRKN, DJ-1, and PINK1 screening identifies novel splice site mutation in PRKN and two novel DJ-1 mutations. Mov. Disord. 26: 80–89. http://onlinelibrary.wiley.com/doi/10.1002/mds.23417/abstract;jsessionid=CC3F0713D4F6E49318C972E74837B2DB.f02t01
- **Hosseini, S.-R.,** M. Sadeghi, H. Pezeshk, C. Eslahchi, and M. Habibi. 2008. PROSIGN: A method for protein secondary structure assignment based on three-dimensional coordinates of consecutive Cα atoms. Comput. Biol. Chem. 32: 406–411. http://www.sciencedirect.com/science/article/pii/S1476927108001072?via=ihub